

# Nettoyage et enrichissement du fichier contestants.csv dans OpenRefine

## Création du projet

Je crée mon dossier, **M2\_DEFI\_2023\_Guerin\_Eurovision**, en important le fichier contestants.csv.

OpenRefine M2\_DEFI\_2023\_Guerin\_Eurovision

Facette / Filtre Défaire / Refaire 0 / 0 1734 lignes

Utiliser les facettes et les filtres

Utiliser les facettes et les filtres pour sélectionner les sous-ensembles de données à traiter. Choisir les méthodes de facette et de filtre dans les menus situés dans les entêtes de colonne.

Vous ne savez pas par où commencer ?  
[Regarder ces tutoriels vidéos](#)

|    | year | to_country_id | to_country  | performer   | song                            | place_contest | sf_num | running_final | running_sf | place_final | points_final | place_sf | points_sf | points_te |
|----|------|---------------|-------------|-------------|---------------------------------|---------------|--------|---------------|------------|-------------|--------------|----------|-----------|-----------|
| 1. | 1956 | ch            | Switzerland | Lys Assia   | Refrain                         | 2             |        | 2             |            | 2           |              |          |           |           |
| 2. | 1956 | nl            | Netherlands | Jetty Paerl | De Vogels Van Holland           | 2             |        | 1             |            | 2           |              |          |           |           |
| 3. | 1956 | be            | Belgium     | Fud Leclerc | Messieurs Les Noyés De La Seine | 2             |        | 3             |            | 2           |              |          |           |           |

Figure 1 Création du projet

## Suppression de colonnes

Suppression de la colonne « to\_country\_id » pour ne se baser que sur « to\_country » dans mon analyse.

OpenRefine M2\_DEFI\_2023\_Guerin\_Eurovision

Remove column to\_country\_id Défaire

Facette / Filtre Défaire / Refaire 1 / 1 1734 lignes

Extraire... Appliquer...

Filter

0. Create project

1. Remove column to\_country\_id

|    | year | to_country  | performer   | song                            | place_contest | sf_num | running_final | running_sf | place_final | points_final | place_sf | points_sf | points_te |
|----|------|-------------|-------------|---------------------------------|---------------|--------|---------------|------------|-------------|--------------|----------|-----------|-----------|
| 1. | 1956 | Switzerland | Lys Assia   | Refrain                         | 2             |        | 2             |            | 2           |              |          |           |           |
| 2. | 1956 | Netherlands | Jetty Paerl | De Vogels Van Holland           | 2             |        | 1             |            | 2           |              |          |           |           |
| 3. | 1956 | Belgium     | Fud Leclerc | Messieurs Les Noyés De La Seine | 2             |        | 3             |            | 2           |              |          |           |           |

Figure 2 Suppression de la colonne to\_country

Suppression des colonnes concernant les demi-finales (sf\_num, running\_sf, place\_sf, points\_sf, points\_tele\_sf, points\_jury\_sf) puisqu'on s'intéresse aux participants des finales (et plus précisément des gagnants). De plus, la France ne participant pas aux demi-finales (membre du Big Five), il n'y a pas d'intérêt.

|                                 | song | place_contest | running_final | place_final | points_final | points_tele_final | points_jury_final | composers               | lyricists            |
|---------------------------------|------|---------------|---------------|-------------|--------------|-------------------|-------------------|-------------------------|----------------------|
| Refrain                         | 2    | 2             | 2             |             |              |                   |                   | Georg Benz Stahl        | (Refrain d'amour...) |
| De Vogels Van Holland           | 2    | 1             | 2             |             |              |                   |                   | Cor Lemaire             | Annie M. G. Schmidt  |
| Messieurs Les Noyés De La Seine | 2    | 3             | 2             |             |              |                   |                   | Jacques Say, Jean Miret | Robert Montal        |

Figure 3 Suppression des colonnes concernant les demi-finales

Supprimer les colonnes composers et lyricists, inutiles pour notre recherche.

|                                 | song | place_contest | running_final | place_final | points_final | points_tele_final | points_jury_final | lyricists                                 |
|---------------------------------|------|---------------|---------------|-------------|--------------|-------------------|-------------------|---|
| Refrain                         | 2    | 2             | 2             |             |              |                   |                   | (Refrain d'amour...)                      |
| De Vogels Van Holland           | 2    | 1             | 2             |             |              |                   |                   | De vogels van Holland zijn zo muzikaal... |
| Messieurs Les Noyés De La Seine | 2    | 3             | 2             |             |              |                   |                   | Messieurs les noyés de la Seine...        |

Figure 4 Suppression des colonnes composers et lyricists

## Enrichissement des pays via WikiData

Je fais correspondre les pays avec l'entité « country » de Wikidata, pour ensuite pouvoir récupérer les données de géolocalisation. Je m'aperçois néanmoins qu'il faut faire souvent le match manuellement car WikiData propose le pays à travers les siècles (par exemple Royaume de France pour la France). Fort heureusement, j'applique le choix pour chaque pays sur plusieurs lignes à chaque fois.

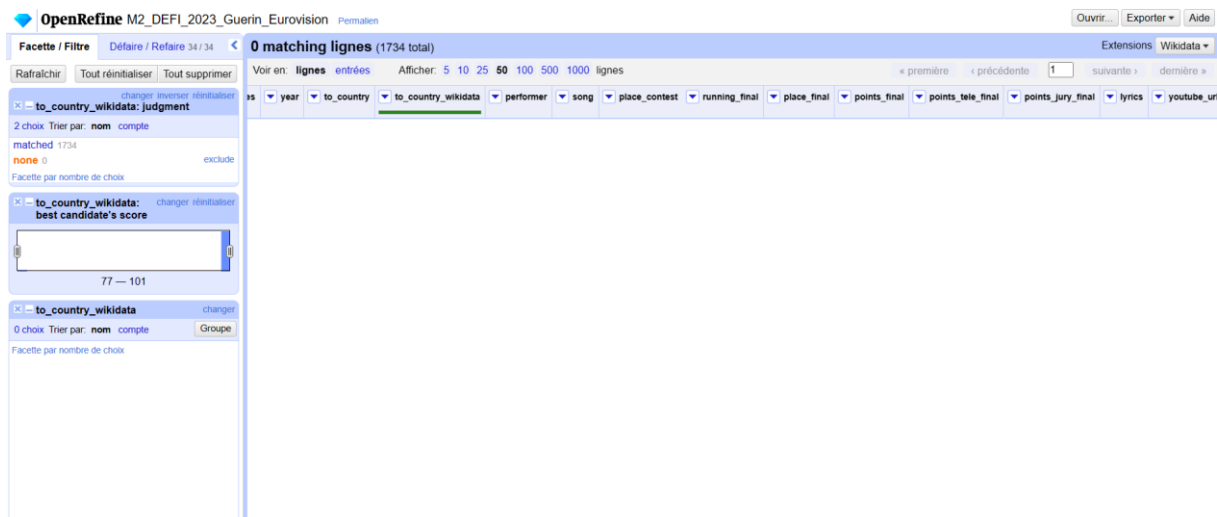


Figure 5 Enrichissement des pays via Wikidata

Ajout de la localisation géographique : pour ce faire, je pars de « to\_country\_wikidata » et je demande à avoir l'entité « coordinate location ».

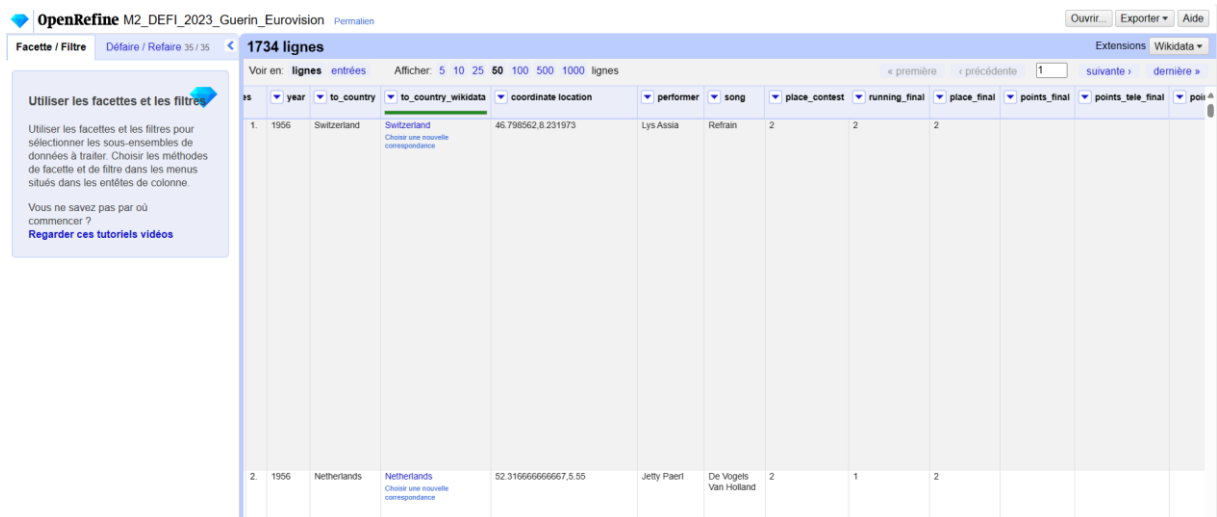


Figure 6 Ajout de la localisation géographique de chaque pays ayant participé à l'Eurovision

## Formatage de l'année

Je crée une colonne « year\_formatUS » pour avoir le bon format pour les outils de visualisation. Je procède de la manière suivante : je crée une colonne (« year\_formatUs ») en me basant sur year, que je transforme en date. Puis, je transforme ces dates de nouveau en texte. Je fais un chercher/remplacer pour enlever le bout d'heure.

| year | year_formatUS | to_country  | to_country_wikidata | coordinate location  | performer   | song                  | place_contest | running_final | place_final | points_final | points |
|------|---------------|-------------|---------------------|----------------------|-------------|-----------------------|---------------|---------------|-------------|--------------|--------|
| 1956 | 1956-01-01    | Switzerland | Switzerland         | 46.796562,8.231973   | Lys Assia   | Refrain               | 2             | 2             | 2           |              |        |
| 1956 | 1956-01-01    | Netherlands | Netherlands         | 52.316666666667,5.55 | Jetty Paerl | De Vogels Van Holland | 2             | 1             | 2           |              |        |

Figure 7 Création d'une colonne au format américain

## Enrichissement des artistes

Je lie les artistes (« performer ») à leur entité « human » de Wikidata. Pour ce faire, je duplique ma colonne « performer » en « performer\_wikidata ». Je crée une colonne « edition » qui est la concaténation de « Eurovision Song Contest » et de la valeur année (year). Le but est d'utiliser cette colonne pour avoir un meilleur taux de match sur la colonne performer\_wikidata puisqu'il existe l'entité « participant in » sur les fiches des artistes. Après cette première opération il y a 589 not matched. Je tente de réconcilier les not match avec musical group : il reste 284 not matched. Puis je tente sur musical duo : il reste 278 not match. Je décide de les laisser pour l'instant de côté.

| location | performer                      | performer_wikidata             | song                   | place_contest | running_final | place_final | points_final | points_tete_final | points_jury_final | lyric  |
|----------|--------------------------------|--------------------------------|------------------------|---------------|---------------|-------------|--------------|-------------------|-------------------|--|
|          | Birthe Wilke & Gustav Winckler | Birthe Wilke & Gustav Winckler | Skibet Skal Seje I Nat | 3             | 9             | 3           | 10           |                   |                   | Farevel d<br>grr l'hor<br>crueller<br>nationV<br>ktsind, t<br>der kuu<br>nationV<br>os them<br>hedrns<br>skal seje<br>kajmriP<br>nu sdr c<br>da, min<br>fin og re<br>skibet s<br>nationE<br>kan jeg i<br>ske nOg<br>skal seje<br>signalk<br>men'ak<br>solidrD<br>skal bost<br>skibet s |
|          | Pearl Carr & Teddy Johnson     | Pearl Carr & Teddy Johnson     | Sing Little Birdie     | 2             | 10            | 2           | 16           |                   |                   | Sing, sir<br>bird on i<br>in the m<br>bird on i<br>tree in t<br>tete bird<br>alongrE<br>kovern<br>treeW<br>seemW<br>eternity<br>our love<br>kovern<br>theAnd t<br>the sdr  |

Figure 8 API WikiData pour les artistes (human, musical duo, musical group)

Pour les match (1456 lignes), je cherche le genre (homme/femme) via Wikidata avec la propriété gender. Une fois ces informations récoltées, je passe ces informations en texte en supprimant le lien avec Wikidata.

Sur ces 1456 lignes d'artistes identifiées, il y a 311 cellules vides. Pour ces cellules vides, je remplis « non applicable » car ce sont les groupes / duos qui restent et qui ne peuvent pas avoir de genre.

| wikidata                         | coordinate location | performer   | performer_wikidata | sex or gender | song                            | place_contest | running_final | place_final | points_final | points_tele_final |
|----------------------------------|---------------------|-------------|--------------------|---------------|---------------------------------|---------------|---------------|-------------|--------------|-------------------|
| 46.798562,8.231973               |                     | Lys Assia   | Lys Assia          | female        | Refrain                         | 2             | 2             | 2           |              |                   |
| 52.3166666666667,5.55            |                     | Jetty Paerl | Jetty Paerl        | female        | De Vogels Van Holland           | 2             | 1             | 2           |              |                   |
| 50.6411111111111,4.6680555555556 |                     | Fud Leclerc | Fud Leclerc        | male          | Messieurs Les Noyés De La Seine | 2             | 3             | 2           |              |                   |

Figure 9 Genre de l'artiste pour les artistes identifiés

Je crée une colonne « performer\_type » à partir de celle concernant le genre. Quand la case indique non applicable, je remplace par groupe. Si elle indique male / female / etc, j'indique solo.

| wikidata | coordinate location | performer                      | performer_wikidata             | sex or gender | performer_type | song                   | place_contest | running_final | place_final | point |
|----------|---------------------|--------------------------------|--------------------------------|---------------|----------------|------------------------|---------------|---------------|-------------|-------|
| 56,10    |                     | Birthe Wilke & Gustav Winckler | Birthe Wilke & Gustav Winckler |               |                | Skibet Skal Seje I Nat | 3             | 9             | 3           | 10    |
| 54,6,-2  |                     | Pearl Carr & Teddy Johnson     | Pearl Carr & Teddy Johnson     |               |                | Sing Little Birdie     | 2             | 10            | 2           | 16    |

Figure 10 Création des colonnes gender et performer\_type

Je crée une colonne test pour les 278 lignes des artistes non trouvés. Ensuite, je fais une recherche textuelle de « & » et « and » pour remplir les cases sur le genre (non applicable : ce sont des groupes) et type de performance en groupe.

OpenRefine M2\_DEFI\_2023\_Guerin\_Eurovision [Permalien](#) **Mass edit 68 cells in column performer\_type** [Défaire](#) [Ouvrir...](#) [Exporter](#) [Aid](#)

Facette / Filtre [Défaire / Refaire 68 / 60](#) **148 matching lignes** (1734 total) Extensions Wikidata

Rafalchir Tout réinitialiser Tout supprimer

Voir en: **lignes** entrées Afficher: 5 10 25 50 100 500 1000 lignes « première » « précédente » 1 « suivante » « dernière »

| performer_type  | wikidata | coordinate location     | performer          | performer_wikidata   | test               | sex or gender | performer_type | song             | place_contest | running_final | place |
|---|----------|-------------------------|--------------------|--|--------------------|---------------|----------------|------------------|---------------|---------------|-------|
| groupe 441<br>solo 1145<br>(blank) 148<br>Facette par nombre de choix |          | 38.7,-9.183333333333334 | Simone de Oliveira | Simone de Oliveira<br><a href="#">Créer un nouvel élément</a><br><a href="#">Chercher une correspondance</a> | Simone de Oliveira |               |                | Sol De Inverno   | 13            | 12            | 13    |
|   |          | 49.77,6.13              | Vicky              | Vicky<br><a href="#">Créer un nouvel élément</a><br><a href="#">Chercher une correspondance</a>              | Vicky              |               |                | L'amour Est Bleu | 4             | 2             | 4     |

Figure 11 Remplissage des groupes en & et and

Je passe à 148 lignes encore incomplètes, que je compléterai après export. Je supprime la colonne « test » et « performer\_wikidata » car le genre et le type d'artistes étaient les seules choses qui m'intéressaient. Je garde finalement la colonne « youtube\_url » car elle va me permettre de voir pour ces 148 lignes directement si c'est un groupe, une femme, un homme etc.

OpenRefine M2\_DEFI\_2023\_Guerin\_Eurovision [Permalien](#) **Mass edit 68 cells in column performer\_type** [Défaire](#) [Ouvrir...](#) [Exporter](#) [Aid](#)

Facette / Filtre [Défaire / Refaire 68 / 60](#) **148 matching lignes** (1734 total) Extensions Wikidata

Rafalchir Tout réinitialiser Tout supprimer

Voir en: **lignes** entrées Afficher: 5 10 25 50 100 500 1000 lignes « première » « précédente » 1 « suivante » « dernière »

| performer_type  | wikidata | coordinate location     | performer          | performer_wikidata   | test               | sex or gender | performer_type | song             | place_contest | running_final | place |
|---|----------|-------------------------|--------------------|--|--------------------|---------------|----------------|------------------|---------------|---------------|-------|
| groupe 441<br>solo 1145<br>(blank) 148<br>Facette par nombre de choix |          | 38.7,-9.183333333333334 | Simone de Oliveira | Simone de Oliveira<br><a href="#">Créer un nouvel élément</a><br><a href="#">Chercher une correspondance</a> | Simone de Oliveira |               |                | Sol De Inverno   | 13            | 12            | 13    |
|   |          | 49.77,6.13              | Vicky              | Vicky<br><a href="#">Créer un nouvel élément</a><br><a href="#">Chercher une correspondance</a>              | Vicky              |               |                | L'amour Est Bleu | 4             | 2             | 4     |

Figure 12 Les 148 artistes qu'il reste à traiter après export

### Autres suppressions de colonne et uniformisation sur place\_contest

Je supprime les colonnes « points\_final », « points\_tele\_final », « points\_jury\_final » car c'est la place obtenue (« place\_contest ») qui m'intéressent. De plus, les évolutions du concours font que ce n'est pas pertinent de prendre ces paramètres en compte.

OpenRefine M2\_DEFI\_2023\_Guerin\_Eurovision

Remove column points\_jury\_final Défaire

Ouvrir... Exporter... Aide

Facette / Filtre Défaire / Refaire 67 / 67 1734 lignes

Extraire... Appliquer...

Extensions Wikidata

Voir en: lignes entrées Afficher: 5 10 25 50 100 500 1000 lignes

« première » précédente 1 suivante » dernière »

|   |    | year | edition                      | year_formatUS | to_country  | to_country_wikidata                                | coordinate location             | performer   | sex or gender | performer_type | song                            | place |
|---|----|------|------------------------------|---------------|-------------|--|---------------------------------|-------------|---------------|----------------|---------------------------------|-------|
| 53. Mass edit 473 cells in column performer_type  | 1. | 1956 | Eurovision Song Contest 1956 | 1956-01-01    | Switzerland | Switzerland<br>Choisir une nouvelle correspondance | 46.798562,8.231973              | Lys Assia   | female        | solo           | Refrain                         | 2     |
| 54. Mass edit 1 cells in column performer_type  | 2. | 1956 | Eurovision Song Contest 1956 | 1956-01-01    | Netherlands | Netherlands<br>Choisir une nouvelle correspondance | 52.316666666667,5.55            | Jetty Paerl | female        | solo           | De Vogels Van Holland           | 2     |
| 55. Create new column test based on column performer_wikidata by filling 278 rows with grel value | 3. | 1956 | Eurovision Song Contest 1956 | 1956-01-01    | Belgium     | Belgium<br>Choisir une nouvelle correspondance     | 50.641111111111,4.6680555555556 | Fud Leclerc | male          | solo           | Messieurs Les Noyés De La Seine | 2     |

56. Mass edit 68 cells in column performer\_type

57. Mass edit 68 cells in column sex or gender

58. Mass edit 62 cells in column performer\_type

59. Mass edit 62 cells in column sex or gender

60. Mass edit 68 cells in column performer\_type

61. Remove column performer\_wikidata

62. Remove column test

63. Remove column running\_final

64. Remove column place\_final

65. Remove column points\_final

Figure 13 Suppression des colonnes

Pour la colonne place\_contest, j'uniformise les 1.0 par exemple en 1. Pour ce faire j'ai cherché dans la colonne tous les « .0 » puis j'ai fait un chercher/remplacer.

OpenRefine M2\_DEFI\_2023\_Guerin\_Eurovision

Ouvrir... Exporter... Aide

Facette / Filtre Défaire / Refaire 68 / 68 1734 lignes

Extraire... Appliquer...

Extensions Wikidata

Voir en: lignes entrées Afficher: 5 10 25 50 100 500 1000 lignes

« première » précédente 1 suivante » dernière »

|   |    | year | edition                      | year_formatUS | to_country  | to_country_wikidata                                | coordinate location             | performer   | sex or gender | performer_type | song                            | place_contest |
|---|----|------|------------------------------|---------------|-------------|--|---------------------------------|-------------|---------------|----------------|---------------------------------|---------------|
| 54. Mass edit 1 cells in column performer_type  | 1. | 1956 | Eurovision Song Contest 1956 | 1956-01-01    | Switzerland | Switzerland<br>Choisir une nouvelle correspondance | 46.798562,8.231973              | Lys Assia   | female        | solo           | Refrain                         | 2             |
| 55. Create new column test based on column performer_wikidata by filling 278 rows with grel value | 2. | 1956 | Eurovision Song Contest 1956 | 1956-01-01    | Netherlands | Netherlands<br>Choisir une nouvelle correspondance | 52.316666666667,5.55            | Jetty Paerl | female        | solo           | De Vogels Van Holland           | 2             |
| 56. Mass edit 68 cells in column performer_type   | 3. | 1956 | Eurovision Song Contest 1956 | 1956-01-01    | Belgium     | Belgium<br>Choisir une nouvelle correspondance     | 50.641111111111,4.6680555555556 | Fud Leclerc | male          | solo           | Messieurs Les Noyés De La Seine | 2             |

57. Mass edit 68 cells in column sex or gender

58. Mass edit 62 cells in column performer\_type

59. Mass edit 62 cells in column sex or gender

60. Mass edit 68 cells in column performer\_type

61. Remove column performer\_wikidata

62. Remove column test

63. Remove column running\_final

64. Remove column place\_final

65. Remove column points\_final

66. Remove column points\_teie\_final

67. Remove column points\_jury\_final

68. Text transform on 37 cells in column place\_contest: value.replace("0","")

Figure 14 Uniformisation de place\_contest

Il reste 56 cellules blanches. En affinant avec l'année, je me rends compte qu'il y a les 40 chansons pour l'année 2020 qui a été annulée. Je me positionne sur l'année 2020 et je mets dans la colonne place\_contest « non applicable ». Il reste 15 chansons pour l'année 2023. Je les remplirai une fois l'export fait.





erlin\_Eurovision [Permalien](#) [Ouvrir...](#) [Exporter](#) [Aide](#)

**1734 lignes** [Extensions](#) [Wikidata](#)

Voir en: **lignes** [entrées](#) Afficher: 5 10 25 50 100 500 1000 lignes « première » « précédente » 1 « suivante » « dernière »

| to_country  | to_country_wikidata  | coordinate location | performer | sex or gender | performer_type | song    | place_contest | lyrics  | langue | youtube_url   |
|-------------|--|---------------------|-----------|---------------|----------------|---------|---------------|---|--------|---|
| Switzerland | <a href="#">Switzerland</a><br><a href="#">Choisir une nouvelle correspondance</a> | 46.798562,8.231973  | Lys Assia | female        | solo           | Refrain | 2             | (Refrain<br>d'amour...)\n\nRefrain,<br>couleur du ciel, parfum<br>de mes vingt<br>ans\nJardin plein de<br>soleil où je courais<br>enfant\nPartout je t'ai<br>cherché, mon amoureux<br>lointain\nCueillant par<br>les sentiers où tu<br>prenais ma main\nLes<br>jours s'en sont allés et<br>nous avons<br>grandi\nL'amour nous a<br>blessés, le temps nous<br>a guéris\nMais seule et<br>sans printemps\nJe<br>cours en vain les bois,<br>les champs\nDis,<br>souviens-toi nos amours<br>d'autrefois?\n\nLes<br>années passent à tire-<br>d'aile\nEt sur les toits de<br>mon ennui coule la<br>pluie\nOù sont parties<br>les caravelles, volant<br>mon cœur\nPortant mes<br>rêves vers ton oubli?<br>\nJ'aurais voulu que tu<br>reviennes comme<br>jadis\nPorter des fleurs<br>à ma persienne\nEt ta<br>jeunesse en mon<br>logis\n\nRefrain, couleur<br>de pluie, regret de mes<br>vingt ans\nChagrin,<br>mélancolie de m'être<br>plus enfant\nMais seule<br>et loin de toi, par les<br>chemins où tu m'es | fr     | <a href="https://youtube.com/watch?v=lyqIPvOkIRk">https://youtube.com/watch?v=lyqIPvOkIRk</a> |

Figure 17 Résultat final dans OpenRefine

Je peux supprimer mon projet ajout\_langue.

Export de l'historique et du fichier

J'exporte mon historique, ainsi que mon fichier au format csv.

Historique : M2\_DEFI\_2023\_Guerin\_Eurovision\_Historique.json

Export au format CSV : M2\_DEFI\_2023\_Guerin\_Eurovision\_ExportOpenRefine.csv

Export au format EXCEL : M2\_DEFI\_2023\_Guerin\_Eurovision\_ExportOpenRefine.xlsx

Traitement Post OpenRefine

Pour rappel je dois compléter :

- 148 lignes pour le genre et le type d'artistes
- 15 places au concours
- 12 chansons avec langues manquantes.

Je duplique mon fichier et je travaille sur ce fichier :

M2\_DEFI\_2023\_Guerin\_Eurovision\_PostOpenRefine.csv

Je commence par les places au concours (place\_contest). Je me rends compte qu'il y a des doublons : Loreen (gagnante de 2023) apparait deux fois. Je pense qu'il y a eu un petit bug (dans le scrapping ou moi en faisant une manipulation). Je supprime du coup mes 12 lignes en trop.

[illegible]

Figure 18 Lignes en double

Pour les 12 chansons sans langue, je vois qu'il y a eu un décalage sur une ligne : un bout des lyrics est dans la case langue et donc, la langue est seule dans 3 colonnes. Je fais les corrections et je supprime les trois colonnes créés (peut-être un souci dans mon script python ?).

Figure 19 Langues manquantes

Il me reste 147 artistes à traiter, que je traite manuellement. Ensuite, il restait deux chansons sans titre. Une fois cette dernière opération, mon fichier est prêt pour utilisation.

M2\_DEF2\_2023\_Guerin\_Eurovision\_PostOpenRefine.csv - LibreOffice

FichierÉditionAffichageInsertionFormatStylesFeuilleDonnéesOutilsFenêtreAide

Libération Sans

10 pt

G I S

</