Projet Image

Social Sciences and Humanities Research Council of Canada | Conseil de recherches en sciences humaines du Canada | Canada

# *Détection de décennies de parution d'une collection de musique par apprentissage*

**Rémy Kessler ,** Nicolas Béchet, Audrey Laplante, Dominic Forest

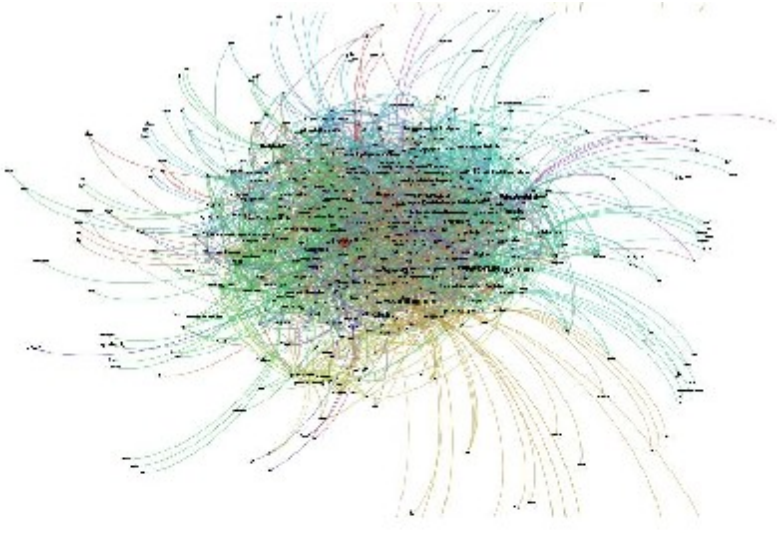
UMR IRISA

Université de Montréal

# *Dans l'épisode précédent...*

Les objectifs du projet IMAGE* étaient :

- représenter une collection de plusieurs milliers de documents musicaux
- pouvoir explorer cette collection en utilisant les paroles comme point d'entrée
- Visualiser les clusters et les mots-clés thématiques associés
- Être capable de filtrer les informations en fonction d'une période musicale spécifique



* : Indexation de Musique A Grande Échelle

# Constitution d'une collection

- Collections existantes
  - « Million Song Dataset » Bertin-Mahieux et al. (2011)
  - « MusiXmatch Dataset » paroles de 200 000 chansons du MSD

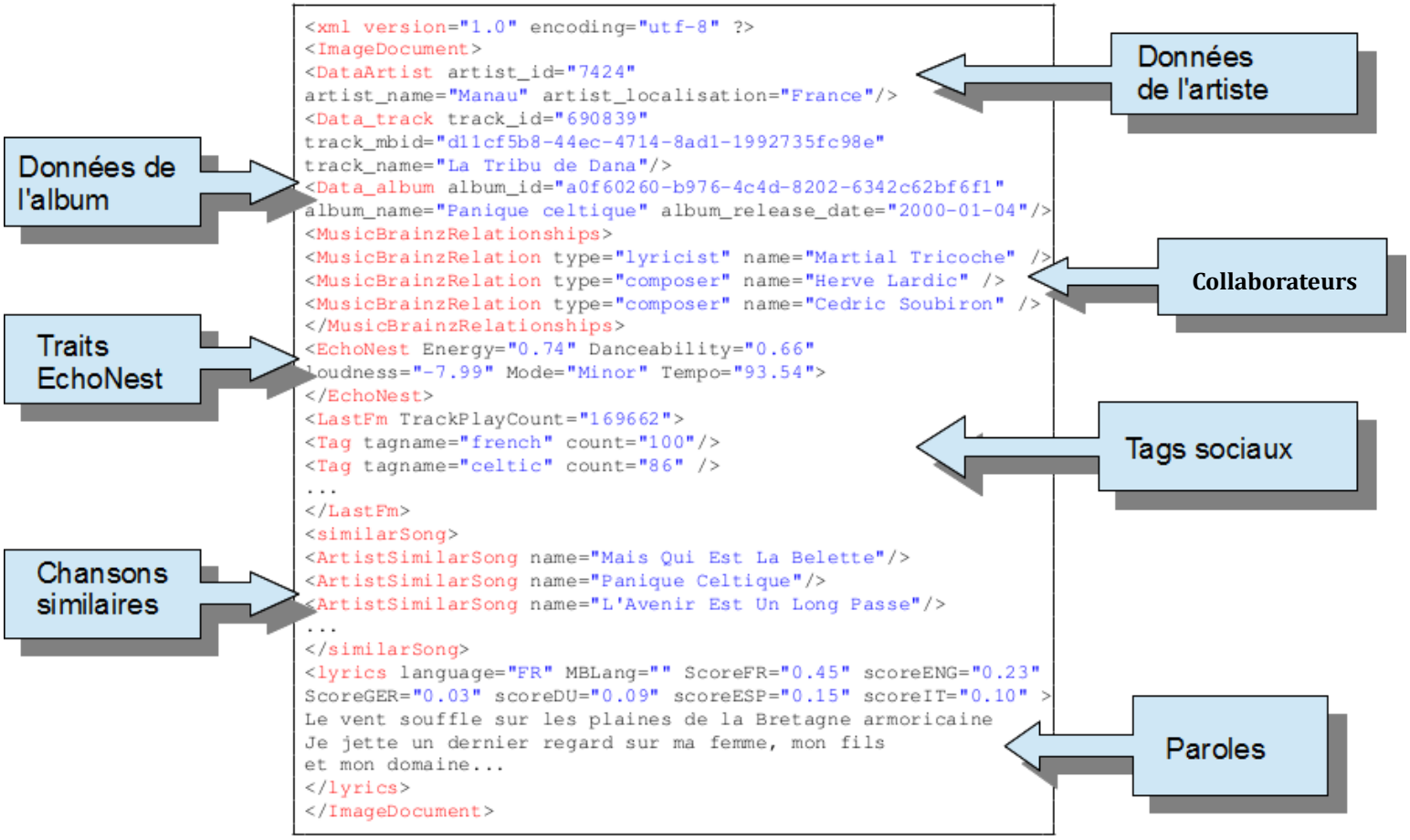
Mise en place d'un système de collecte


Statistiques de collecte avant les filtrages
- 45 000 en français
- 53 000 autres langues

# *statistiques des collections*

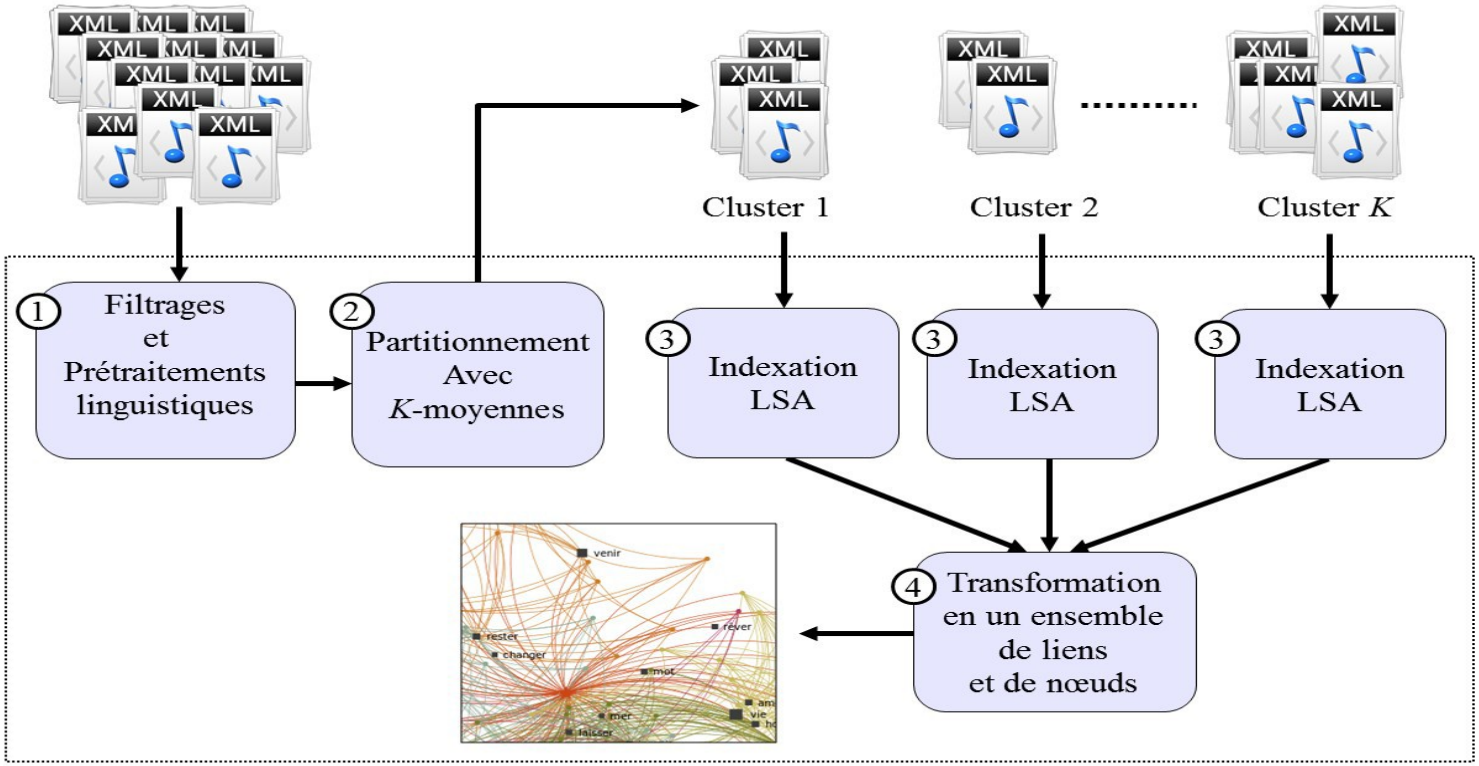
| | Collection | |
|---|---|---|
| Nombre total de chansons | 4 529 | 12 109 |
| ***Sans pré-traitement linguistique*** | | |
| Nombre total de mots | 1 117 951 | 2 803 194 |
| Nombre total de mots différents | 74 311 | 131 727 |
| Moyenne de mots par chanson | 239,21 | 227,77 |
| ***Avec pré-traitement linguistique*** | | |
| Nombre total de mots | 482 609 | 1 217 965 |
| Nombre total de mots différents | 31 305 | 57 341 |
| Moyenne de mots par chanson | 105,57 | 101,28 |

# Exemple de document

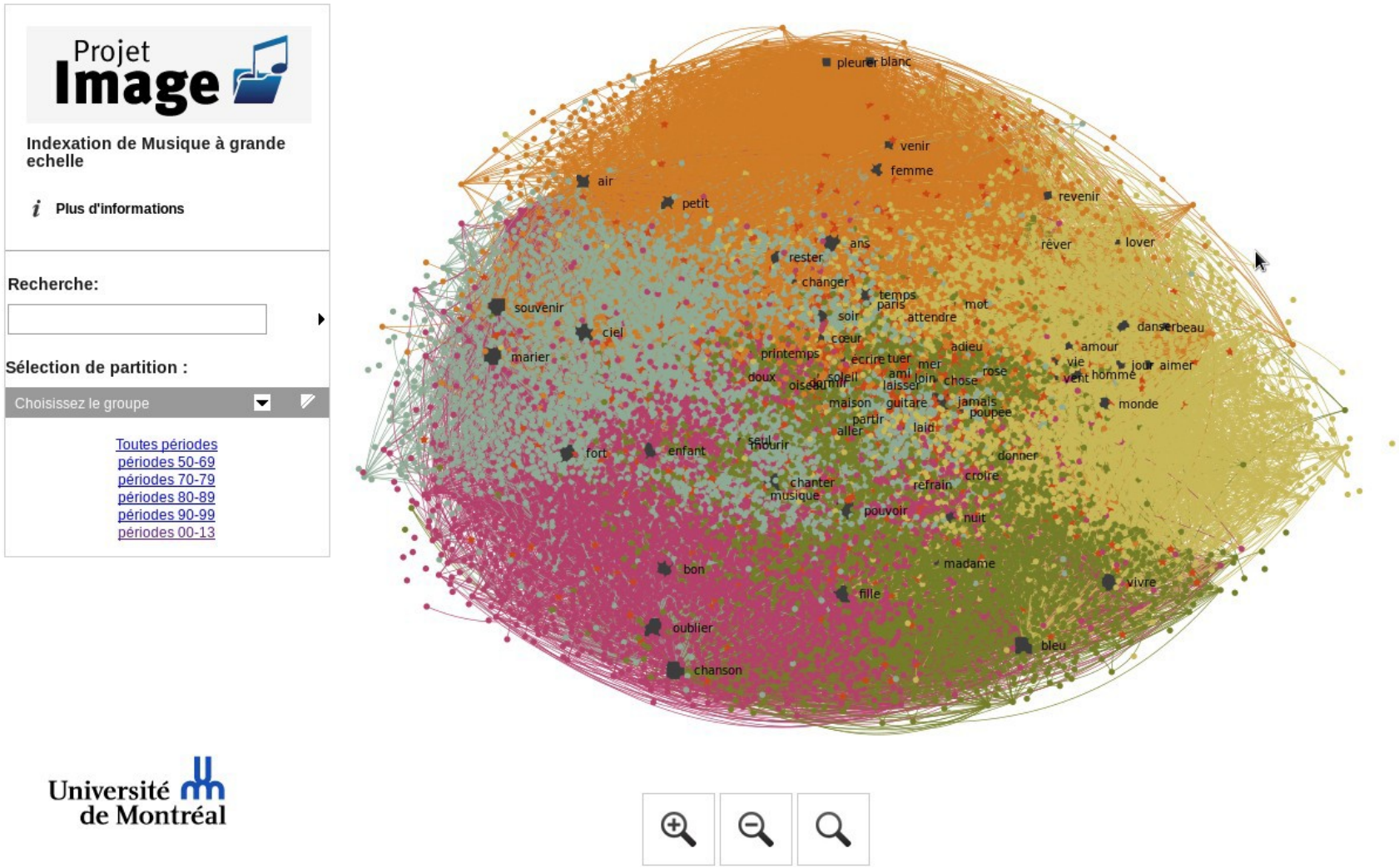



```
<xml version="1.0" encoding="utf-8" ?>
<ImageDocument>
<DataArtist artist_id="7424"
artist_name="Manau" artist_localisation="France"/>
<Data_track track_id="690839"
track_mbid="d11cf5b8-44ec-4714-8ad1-1992735fc98e"
track_name="La Tribu de Dana"/>
<Data_album album_id="a0f60260-b976-4c4d-8202-6342c62bf6f1"
album_name="Panique celtique" album_release_date="2000-01-04"/>
<MusicBrainzRelationships>
<MusicBrainzRelation type="lyricist" name="Martial Tricoche" />
<MusicBrainzRelation type="composer" name="Herve Lardic" />
<MusicBrainzRelation type="composer" name="Cedric Soubiron" />
</MusicBrainzRelationships>
<EchoNest Energy="0.74" Danceability="0.66"
Loudness="-7.99" Mode="Minor" Tempo="93.54">
</EchoNest>
<LastFm TrackPlayCount="169662">
<Tag tagname="french" count="100"/>
<Tag tagname="celtic" count="86" />
...
</LastFm>
<similarSong>
<ArtistSimilarSong name="Mais Qui Est La Belette"/>
<ArtistSimilarSong name="Panique Celtique"/>
<ArtistSimilarSong name="L'Avenir Est Un Long Passe"/>
...
</similarSong>
<lyrics language="FR" MBLang="" ScoreFR="0.45" scoreENG="0.23"
ScoreGER="0.03" scoreDU="0.09" scoreESP="0.15" scoreIT="0.10" >
Le vent souffle sur les plaines de la Bretagne armoricaine
Je jette un dernier regard sur ma femme, mon fils
et mon domaine...
</lyrics>
</ImageDocument>
```

# Traitement des données



# *IMAGE en image...*



# *IMAGE en vidéo…*

# *mots-clés thématiques pour les 5 clusters*

| | |
|---|---|
| Cluster 1 (mandarine) | aimer, amour, temps, jour, dire, vie, beau |
| Cluster 2 (jaune) | falloir, fille, monde, homme, nuit, soir, bleu |
| Cluster 3 (framboise) | croire, petit, mer, partir, revenir, souvenir, loin |
| Cluster 4 (vert foncé) | venir, nord, tourner, lune, danser, bon, pleurer |
| Cluster 5 (aqua) | vivre, mal, joie, falloir, peiner, attendre, perdre |

**Interprétation**

Cluster 3 : thèmes du voyage ou des départs
Cluster 2 : chansons d'amour tristes
Cluster 1 : chansons d'amour positives

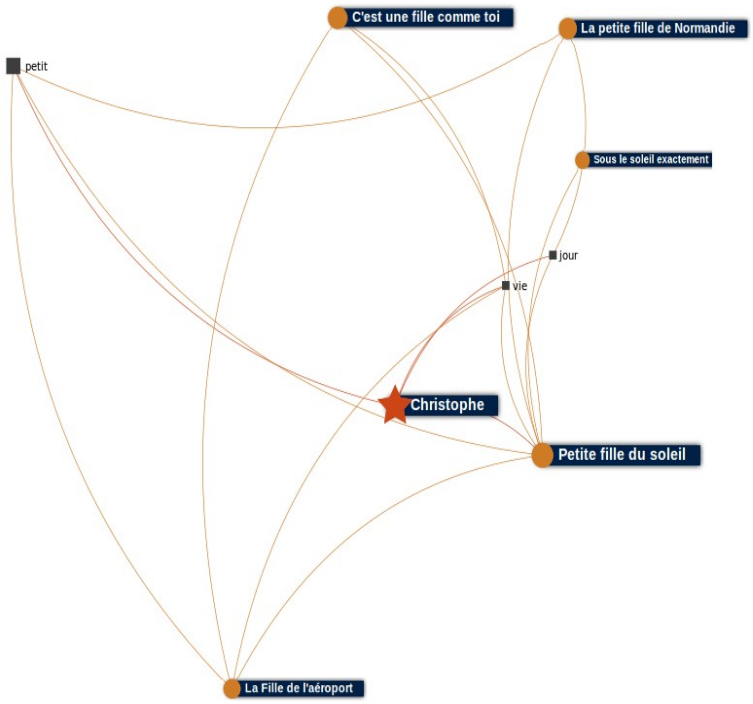
# *répartition des clusters en fonction des périodes musicales*

| | Cluster | | | | |
|---|---|---|---|---|---|
| Périodes | 1 | 2 | 3 | 4 | 5 |
| Années 50-69 | 17,28% | 24.28% | 21,81% | 18,93% | 17,70% |
| Années 70-79 | 22,47% | 19,59% | 20,41% | 19,79% | 17,73% |
| Années 80-89 | 20,65% | 19,80% | 20,86% | 19,66% | 19,03% |
| Années 90-99 | 20,35% | 20,39% | 19,42% | 20,04% | 19,80% |
| Années 2000-2013 | 20,26% | 20,38% | 18,93% | 20,16% | 20,26% |

# Problèmes

- Période manquante pour une partie de la collection

- Compléter les informations contenues dans la collection en évitant une tâche fastidieuse d'étiquetage manuel




*Sous-réseau de la chanson « Petite fille du soleil ».*

# *Méthodologie*

- Système de base évalué lors de la campagne DEFT 2011 (Oger et al., 2010)
- Représentation sous forme n grammes de mots
- Approche par classification multi-classe (5 classes différentes, 1 classe par période)
- Les paramètres d'entrées du système sont les entrées lexicales de chaque document (les paroles des chansons)
- Traits supplémentaires :
  - Traits EchoNest : valeurs numériques pour chaque chanson transmise par le site EchoNest (ex : «danceability», «Hotness», «loudness» )
  - Collaborateurs : listes des collaborateurs (parolier, compositeur, arrangeur, musiciens, etc.)
  - Chansons et artistes similaires : Ensemble des chansons et artistes similaires selon les informations issues des sites *LastFM* et *EchoNest*.
  - Tags sociaux : tags sociaux du site Last.fm associés à chaque chanson

# *motifs séquentiels*

- Chaque chanson est représentée comme une base de données séquentielles (SDB)
  Exemple : "Sauver une âme" devient :
  <(Sauver VERBE) (un PRONOM ) (âme NOM)>

- Méthode :

  1) Construire une SDB pour chaque chanson
  2) Extraire les motifs à partir des SDB
  3) Conserver les k motifs les plus fréquents

- Mise en œuvre d'une mesure de similarité afin de comparer les chansons pour calculer les médoïdes qui représenteront les classes.

- Résultats préliminaires *(exemple de motifs extraits)*:

{**NOM_**}{**NOM_** personne} sup=**4**
{**NOM_**}{**ADV_**} sup=**5**
{**VER**:pres_}{**ADV_**} sup=**3**
{**ADV_**}{**NOM_**} sup=**5**
{**DET**:**ART_** le}{**NOM_**}{**VER**:pres_} sup=**6**
{**PRP_**}{**DET**:**ART_** le}{**NOM_**} sup=**3**
{**PRP_** de}{**NOM_**}{**NOM_**} sup=**4**
{**PRP_** de}{**ADV_**} sup=**4**

{**PRP_** à}{**NOM_** ombre} sup=**6**
{mon **DET**:**POS_**}{NOM_} sup=**3**
{on **PRO**:**PER_**}{**VER**:pres_}{**NOM_**} sup=**4**
{on **PRO**:**PER_**}{**VER**:pres_}{**PRP_** de} sup=**4**
{**KON_**}{**NOM_**} sup=**5**
{**KON_**}{**ADV_**} sup=**3**
{et **KON_**}{on **PRO**:**PER_**}{**PRO**:**PER_**se}{**VER**:pres_} sup=**4**

# Résultats

| | Précision | Rappel | F-score | Mesure DEFT 2011 |
|---|---|---|---|---|
| **Sans prétraitements linguistiques** | | | | |
| Système de base (SB) | 0,383 | 0,974 | 0,550 | 0,279 |
| SB+traits EchoNest (TE) | 0,388 | 0,954 | 0,551 | 0,282 |
| SB+Collaborateurs (COL) | 0,387 | 0,961 | 0,551 | 0,281 |
| SB+Tags sociaux (TS) | **0,518** | **0,759** | **0,613** | **0,284** |
| SB+Artistes/Chansons similaires (ACS) | 0,383 | 0,974 | 0,550 | 0,278 |
| SB+TE+TS | 0,533 | 0,752 | 0,620 | 0,280 |
| SB+TE+TS+COL | 0,511 | 0,773 | 0,615 | 0,280 |
| SB+TE+TS+COL+ACS | **0,511** | **0,773** | **0,615** | **0,281** |
| **Avec prétraitements linguistiques** | | | | |
| Système de base (SB) | 0,383 | 0,969 | 0,549 | 0,278 |
| SB+traits EchoNest (TE) | 0,388 | 0,971 | 0,554 | 0,282 |
| SB+Collaborateurs (COL) | 0,388 | 0,959 | 0,553 | 0,281 |
| SB+Tags sociaux (TS) | **0,541** | **0,758** | **0,629** | **0,284** |
| SB+Artistes/Chansons similaires (ACS) | 0,383 | 0,969 | 0,549 | 0,278 |
| SB+TE+TS | 0,533 | 0,752 | 0,620 | 0,280 |
| SB+TE+TS+COL | 0,404 | 0,899 | 0,556 | 0,277 |
| SB+TE+TS+COL+ACS | **0,545** | **0,754** | **0,630** | **0,290** |

# *Analyse des résultats*

Influence de certaine informations :

- Artistes et collaborateurs
  - artistes avec des périodes musicales plus grandes
    (par ex : Céline Dion, Johnny Halliday, etc… )
  + compositeur extrêmement productif
    (par ex: Jean Jacques Goldman, Pierre Lapointe)
- Tags sociaux

- « Star », « live », etc.  motifs répétés
+ « seen in 2013 », « 1960s » , « live in 2010 »

# Conclusions

- Site internet afin d'explorer une collection de documents musicaux
- Approche combinant fouille de texte et apprentissage afin de retrouver les périodes musicales manquantes d'une large collection
- L'utilisation des paroles et des tags sociaux sont une source intéressante afin d'identifier les liens entre les différentes périodes musicales.
- Les premiers tests préliminaires basés sur les motifs séquentiels présentent des résultats encourageants

# Perspectives

- Exploration en prenant les tags sociaux comme point d'entrée.
- Détection de parolier, de styles musicaux propres à chaque artiste.

# Références


- Bastian M., Heymann S. and Jacomy M. (2009). *Gephi: An open source software for exploring and manipulating networks*, in International AAAI Conference on Weblogs and Social Media.

- Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. and Harshman R. (1990). *Indexing by latent semantic analysis.* J. Am. Soc. Inf. Sci. 41, pp. 391–407.

- **Rémy Kessler**, Laplante Audrey and Forest Dominic. *Une interface visuelle pour l'exploration d'une collection de musique*. In 14èmes Journées Extraction et Gestion des Connaissances (EGC 2014), Rennes.

- **Rémy Kessler**, Audrey Laplante, Dominic Forest, "Encore des mots, toujours des mots : fouille de textes et visualisation de l'information pour l'exploration et l'analyse d'une collection de chansons en français." 15èmes journées internationales d'analyse statistique des données textuelles JADT 2014.

- **Rémy Kessler**, Nicolas Béchet, Audrey Laplante,Pierre Francois Marteau, Dominic Forest, "Enrichissement d'une collection de musique par apprentissage " 2014 Traitement Automatique de la Langue Naturelle (TALN 2014).

- McKay C., Burgoyne J.A., Hockman J., Smith J.B., Vigliensoni G. and Fujinaga I. (2010). *Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features.* In ISMIR, pp. 213-218.

- Pampalk E. (2001). *Islands of music: Analysis, organization, and visualization of music archives.* Masters thesis Vienna Univ. Technol. *Vienna Austria.*