Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

LaTTiCe

Laboratoire Langues, Textes, Traitements informatiques, Cognition

---

Filtrage sémantique et visualisation de données textuelles

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue

par

Johan FERGUTH

Directeurs de mémoire :

Thierry POIBEAU

Clément PLANCQ

Année universitaire 2014/2015

MÉMOIRE

Résumé

Ce travail porte sur la question de la visualisation thématique en recherche d'informations.
Dans un contexte de plus en plus prégnant de circulation d'informations et face à d'importants flux de données il convient de synthétiser l'information. A plus forte raison garantir un accès rapide et pertinent à l'information peut devenir délicat lorsque c'est un utilisateur qui définit le thème recherché.
Nous proposons une approche par croisement de thèmes "simples" pour définir une thématique "complexe". Nous abordons ensuite un système d'enrichissement thématique basé sur des coefficients de similarité. Enfin nous traitons de la visualisation de données en nous appuyant sur les entités nommées contenues dans la thématique détectée .
On considérera ici qu'un utilisateur peut obtenir une réponse à une information recherchée grâce à la synthèse visuelle d'entités nommées issues de la détection de thème.

Mots-clefs : Détection de thème, Visualisation thématique, Entités nommées, Recherche d'information, Visualisation de données

Keywords : Topic detection , Topic visualization, Named entity, Information retrieval, Data visualisation

TABLE DES MATIÈRES

Liste des figures

Liste des tableaux

INTRODUCTION

Contexte

C'est dans les années 90 que l'on voit l'extension et la démocratisation d'internet, et avec eux deviennent accessibles de véritables flots de données [1]. Par extension, l'accessibilité de ces données en traitement automatique des langues, ou *TAL*, se traduit par un renforcement d'approches à tendance empirique. Ces visées applicatives se heurtent néanmoins à la qualité des informations que l'on peut extraire. En effet, assurer et optimiser cette qualité [2] fait souvent figure de graal à atteindre surtout quand l'information recherchée est difficile à déterminer (notamment par une requête). Dans ce contexte, différentes approches voient le jour pour analyser et synthétiser différentes informations.

On citera par exemple les travaux récents de Xavier Tannier qui développe un logiciel « permettant d'analyser, à partir d'articles de presse et de tweets, les opinions des différents courants politiques et de leurs membres » [3]. On peut aussi citer l'analyseur de dépêche du site EMM [4], qui produit une synthèse d'événements en temps réel.

Dans le cadre de ce mémoire, nous nous situons aussi dans cette optique en cherchant à produire une chaîne de traitement de recherche d'informations. Cependant comme nous l'expliquions ci-dessus, face à un gros volume de données, l'accès aux articles pertinents semble ne plus suffire. À l'image des travaux de Tannier nous souhaiterions produire une synthèse d'informations sous forme de visualisations (ou *dataviz*) de façon à explorer les résultats de la recherche d'information. L'aspect innovant de ce travail s'appuiera sur la notion de thématique, et plus particulièrement sur le croisement de thèmes que l'utilisateur pourra définir à son gré et selon ses besoins. Cette chaîne de traitement se devra donc de répondre à la question suivante :

Comment garantir un accès à une synthèse visuelle pertinente pour un utilisateur, en partant d'un thème défini par ce dernier ?

1. On voit ainsi apparaître pour la première fois le terme ***Big Data*** dans un article de Michael Cox et David Ellsworth [Cox and Ellsworth, 1997], article aussi disponible sous format numérique http://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf. Ce terme sera employé dans le sens qu'on lui connaît aujourd'hui à partir des années 2000, avec l'émergence des réseaux sociaux et l'utilisation des données en contexte économique.

2. Qui correspond grossièrement au ratio : $\frac{pertinence\ des\ informations}{volume\ de\ données}$.

3. Article du journal CNRS https://lejournal.cnrs.fr/articles/un-logiciel-qui-decrypte-la-politique, dernière consultation le 04/11/15.

4. http://emm.newsbrief.eu/NewsBrief/clusteredition/fr/latest.html, dernière consultation le 04/11/15.

Réalisation

Cadre du stage

Le traitement décrit dans ce mémoire s'appuie sur un stage effectué lors d'une collaboration conjointe entre le laboratoire de recherche LaTTiCe [5] sous la direction de M. Thierry Poibeau d'une part, et l'institut national des langues et civilisations orientales (INALCO) [6] sous la direction de M. Clément Plancq d'autre part. Ce stage a de plus fait l'objet d'un financement par le labex EFL (Fondements Empiriques de la Linguistique) [7]. Nous n'avons pu effectuer tous les développements que nous souhaitions lors de ce stage de six mois : ce mémoire est aussi l'occasion d'une réflexion plus large sur les limites et les perspectives de notre travail.

Cadre préparatoire

Voici notre cadre applicatif, il permettra de déterminer les développements à effectuer :

On imagine qu'un expert (par exemple un veilleur dans une entreprise de défense ou un journaliste du Monde) possède un corpus ayant les caractéristiques suivantes :

— Volumineux par l'ampleur des données : avec un volume rendant inabordable la lecture pour un expert humain (ou même un groupe d'experts),
— Dont les données sont faiblement structurées,
— Avec un contenu non homogène, : abordant un nombre de thématiques variées
— Le style, quant à lui, est assez régulier dans la mesure où il s'agit de textes journalistiques [8]

L'expert souhaite explorer les données selon des critères qui lui sont propres, nous parlerons dans notre cas de **thématique** [9]. Nous posons comme hypothèse de travail que la recherche d'informations pour un utilisateur s'effectue fréquemment par regroupements thématiques complémentaires. On souhaite donc qu'il puisse établir des requêtes croisées. Nous tenterons par exemple d'apporter des réponses, par le biais de visualisations, aux requêtes suivantes [10] :

— « *terrorisme* » + « *patrimoine de l'humanité* »
— « *terrorisme* » + « *drone* ».

Ici, si « drone » peut faire l'objet d'une recherche assez efficace dans la mesure où le mot constitue un mot clé peu ambigu, il n'en va pas de même pour « patrimoine de l'humanité ». On est là bien dans le cadre d'un thème : définir « patrimoine de l'humanité » nécessite de trouver un ensemble de mots clés pertinents afin de retrouver des documents ne comprenant finalement quasiment jamais le terme « patrimoine de l'humanité ». Les graphiques se prêtent par essence aux synthèses, une réponse à la problématique pourrait donc se caractériser par une série de visualisations permettant des accès variés aux données, synthétisant l'information recherchée par l'utilisateur de différentes manières. Il apparaîtra que plusieurs outils permettent de résoudre certaines étapes du problème, on imagine ainsi pouvoir créer une chaîne de

5. http://www.lattice.cnrs.fr/
6. http://www.inalco.fr/
7. http://www.labex-efl.org/
8. *cf infra* chapitre Corpus.
9. Nous abordons ici volontairement cette notion en surface pour y revenir plus tard *cf* chapitre Méthodes.
10. Voir le chapitre 3 p.63 concernant les résultats.

traitement, en partant du corpus brut jusqu'aux visualisations en tirant parti de ces outils.

Cette chaîne de traitement devant répondre à des besoins de l'utilisateur, nous tentons donc de les pré-définir ainsi que d'évaluer les réponses à ces problèmes :

Besoins utilisateurs	Approches pour répondre à ces besoins
Établir un accès à l'information recherchée par le biais de requêtes.	La première rencontre avec ce point se fera par le biais de mots clefs (par exemple : terroriste, drone).
Répondre à des requêtes thématiques croisées.	Nous verrons ici comment nous traitons la notion de thématique et comment nous avons choisi de croiser ces thématiques.
Enrichissement des requêtes en s'appuyant sur la proximité thématique	Nous viserons ici un enrichissement par coefficient de similarité comme pour la requête « *patrimoine de l'humanité* + *terrorisme* ».
Visualiser les acteurs (au sens large, par exemple personnes, organisations) de ces thématiques.	Pour ce point nous utiliserons essentiellement comme pivot la notion d'entités nommées incluses dans nos thématiques.
Visualiser les acteurs de ses thématiques croisées, associés à des informations extérieures au corpus.	En cherchant à isoler certaines entités nommées, à les classifier *a priori* ou *a posteriori*, et à les filtrer selon les besoins qui seront les nôtres pour garantir une visualisation adéquate. Nous envisagerons ici un filtrage grâce à des connaissances extérieures (via par exemple des ontologies présentes sur dbpedia) pour accroître les informations présentées.
Visualiser ces mêmes acteurs selon des aspects spatiaux et/ou temporels.	On vise ici une association entre différents types d'entités nommées, avec les classements et filtrages précédents. On espère, en terme de résultats, des représentations de type : — acteurs (au sens effectue ou subit une action) + lieu de l'action — timeline + acteurs.
Accéder aux actions antérieures des mêmes acteurs.	On cherchera à établir un lien entre l'acteur d'un texte et le même acteur plus tôt ou plus tard dans le corpus.
Établir que deux acteurs ont un rapport en commun.	On cherchera à établir des connexions entre acteurs en se basant sur les cooccurrences et en les associant à des patrons syntaxiques.

TABLE 0.1 – Définitions des besoins utilisateurs et approches pour y répondre

Ce cadre fixé nous abordons maintenant la question du plan.

Plan de lecture

Dans un premier temps, avant d'effectuer des traitements plus complexes, nous avons besoin de standardiser les données pour les extraire. Nous expliquerons ce besoin en abordant la question du corpus, les détails concernant sa volumétrie et sa composition. Il sera aussi question des pré-traitements effectués, nécessaires à la préparation et à l'uniformisation du corpus. Enfin n'étant pas nous même expert

d'un domaine, nous utiliserons des outils d'exploration qui nous aiderons à définir notre thématique principale.

Nous regarderons ensuite, dans une deuxième partie, en quoi nous pouvons caractériser un thème et en quoi nous pouvons utiliser cette notion pour qu'un utilisateur puisse interroger des données. Cette caractérisation thématique sera l'occasion d'une tentative de recherche de thèmes similaires (coefficient de similarité).

Cette « définition » du thème nous poussera à identifier les éléments contenus dans l'information recherchée par l'utilisateur qui seront les plus marqués sémantiquement, les plus « faciles » à extraire et, par conséquent les plus visualisables. Conjuguant qualité de l'information et facilité d'extraction, c'est donc naturellement que nous nous tournerons vers les entités nommées. Nous aborderons enfin la question de la visualisation et du travail préparatoire nécessaire à celle-ci.

Dans le chapitre trois nous traiterons de la question des résultats en observant l'utilisation de l'outil constitué sur deux requêtes, comme nous l'avons déjà mentionné (« *terrorisme* » + « *patrimoine de l'humanité* », « *terrorisme* » + « *drone* »).

Enfin nous tenterons d'évaluer et observerons les limites du travail réalisé dans une quatrième partie. Avant d'aborder les perspectives possibles en conclusion.

Nous observons ici, que l'état de l'art se caractérisant, dans notre cas, comme une somme d'informations relatives à chacune des sous-tâches de notre traitement, il convenait de le présenter dans les parties concernées. Nous avons donc fait le choix d'introduire les notions nécessaires à la compréhension de façon localisée, plutôt que globalement.

CHAPITRE 1

CORPUS

Sommaire

1.1 Introduction

Ce chapitre sera l'occasion de présenter notre corpus avec sa volumétrie, et les raisons de sa standardisation. En effet, nous avions mentionné dans l'introduction générale qu'avant d'aborder les traitements sémantiques, plus profonds des données, il convenait de standardiser en surface ces dernières. Cette uniformisation nous donnera un premier accès aux données et sera l'occasion de simuler le regard de l'expert utilisateur grâce à des outils d'exploration textuelle.

1.2 Présentation du Corpus

Nous commençons ici par une première présentation des données. Le corpus que nous utilisons ici est issu d'un script de *crawling*, implémenté par Serge Fleury, effectué sur la page du flux RSS du journal *Le Monde* [1]. Nous rappelons qu'à ce titre, son statut juridique ne permet qu'une utilisation en interne, dans un objectif de recherche, sans publication. Enfin, les éventuelles démonstrations de l'application ne s'effectueront pas en accès ouvert.

La composition du corpus s'étend sur une période de 74 mois, de l'année 2009 à février 2015 [2]. Le résultat de l'aspiration du site internet produit des dossiers organisés en sous-dossiers de la façon suivante :

1. On pourra consulter : http://www.lemonde.fr/ et http://www.lemonde.fr/rss/.
2. Excepté une dizaine de jours manquants.

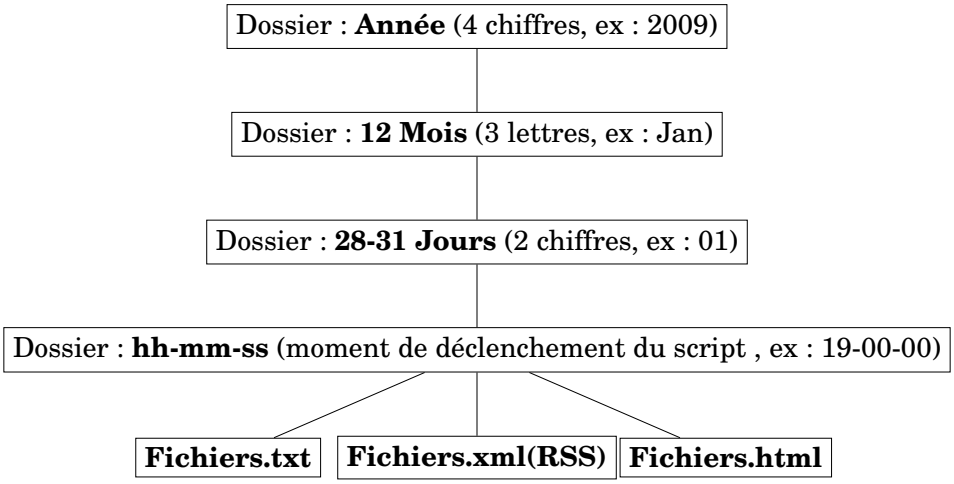


FIGURE 1.1 – Organisation du corpus

Pour chaque dossier, le volume de données avant normalisation correspond à :

Dossier année	**Poids(Go)**
2009	1,7
2010	1,3
2011	1,5
2012	1,6
2013	1,9
2014	2,2
2015 (Jan-Fév)	287,8 (Mo)

TABLE 1.1 – Volume du corpus avant normalisation

Comme nous allons le voir dans la section suivante, en observant les données, on remarque leur faible structuration. Ainsi des éléments externes (comme différents encodages) ou interne aux données (comme le fait que les titres et/ou articles ne s'agencent pas toujours de la même façon) implique de devoir les uniformiser. De cette façon on facilitera les traitements ultérieurs en standardisant les éléments qui peuvent l'être ou en excluant les autres. On garantit aussi que chaque article sera traité dans les mêmes conditions. Nous décrivons ci-dessous le schéma des différentes opérations de cette standardisation. Ce sera aussi l'occasion de quantifier de façon plus précise les données.

1.3 Opérations de standardisation

Dans cette étape de traitement on cherche à uniformiser le corpus. Un bref examen de celui-ci nous permet de constater des différences entre types de format de fichiers.

Les fichiers XML, sont des fils RSS pauvres en contenu. Si les titres sont identiques à ceux des autres fichiers, l'élément *description* ne contient qu'un bref synopsis de l'article. Le reste du fichier se décompose en éléments relatifs aux méta-données repartis en balises selon les spécifications RSS [3] comme dans l'exemple suivant (en

3. Voir spécifications RSS2.0 http://cyber.law.harvard.edu/rss/rss.html, dernière consultation du 02/11/15.

couleur le contenu textuel) :

```
politique/son/2009/04/01/villepin-compte-ses-troupes-a-l-assemblee-nat
ionale_1175460_823448.html#xtor=RSS-3208</guid></item><item><title>L'A
fghanistan prévoit une loi liberticide pour les femmes</title><link>ht
tp://www.lemonde.fr/asie-pacifique/article/2009/04/01/l-afghanistan-pr
evoit-une-loi-liberticide-pour-les-femmes_1175377_3216.html#xtor=RSS-3
208</link><description>Selon "The Guardian", le texte signé par le pré
sident Hamid Karzaï, mais pas encore publié, légalise le viol de sa co
njointe et interdit aux femmes de sortir sans la permission de leur ma
ri.&lt;img width='1' height='1' src='http://rss.feedsportal.com/c/205/
f/3050/s/3a42323/mf.gif' border='0'/&gt;&lt;br/&gt;&lt;br/&gt;&lt;a hr
ef="http://da.feedsportal.com/r/35757345415/u/6/f/3050/c/205/s/6108854
7/a2.htm"&gt;&lt;img src="http://da.feedsportal.com/r/35757345415/u/6/
f/3050/c/205/s/61088547/a2.img" border="0"/&gt;&lt;/a&gt;</description
><enclosure url="http://medias.lemonde.fr/mmpub/edt/ill/2009/04/01/h_1
_ill_836748_afghanistan-femmes.jpg" length="2531" type="image/jpeg" />
<pubDate>Wed, 01 Apr 2009 15:40:19 GMT</pubDate><guid isPermaLink="fal
se">http://www.lemonde.fr/asie-
```

Les fichiers HTML sont tout aussi pauvres en contenu ils ne comportent (d'un point de vue textuel) que titres et descriptions à l'image des fils RSS :

```
<b>Titre</b> : L'Afghanistan prévoit une loi liberticide pour les femm
es<br/> <br/><b>Lien</b> : <a target="blank_" href="http://www.le
monde.fr/asie-pacifique/article/2009/04/01/l-afghanistan-prevoit-une-l
oi-liberticide-pour-les-femmes_1175377_3216.html#xtor=RSS-3208">http:/
/www.lemonde.fr/asie-pacifique/article/2009/04/01/l-afghanistan-prevoi
t-une-loi-liberticide-pour-les-femmes_1175377_3216.html#xtor=RSS-3208<
/a><br/> <br/><b>Description</b> : <br/>Selon "The Guardian", le
texte signé par le président Hamid Karzaï, mais pas encore publié, lég
alise le viol de sa conjointe et interdit aux femmes de sortir sans la
 permission de leur mari.<img width='1' height='1' src='http://rss.fee
dsportal.com/c/205/f/3050/s/3a42323/mf.gif' border='0'/><br/><br/><a h
ref="http://da.feedsportal.com/r/35757345415/u/6/f/3050/c/205/s/610885
47/a2.htm"><img src="http://da.feedsportal.com/r/35757345415/u/6/f/305
0/c/205/s/61088547/a2.img" border="0"/></a><br/> <br/><b>Date de
publication</b> : Wed, 01 Apr 2009 15:40:19 GMT<br/> <br/><b>Lien
 permanent</b> : <a target="blank_" href="http://www.lemonde.fr/asie-p
acifique/article/2009/04/01/l-afghanistan-prevoit-une-loi-liberticide-
pour-les-femmes_1175377_3216.html#xtor=RSS-3208">http://www.lemonde.fr
/asie-pacifique/article/2009/04/01/l-afghanistan-prevoit-une-loi-liber
ticide-pour-les-femmes_1175377_3216.html#xtor=RSS-3208</a><br/> <
br/><p><hr/></p>
```

Les fichiers TXT sont les plus riches en contenu. Ils contiennent également le titre et la date de publication de chaque article [4] :

4. Pour des facilités d'affichage en verbatim LaTeX nous avons supprimé les saut de lignes des fichiers originaux.

<article-nb="2009/04/01/19-2"><filnamedate="20090401"><AAMM="200904"><AAMMJJ="20090401"><AAMMJJHH="2009040119"><filname="SURF-0,2-3208,1-0,0-2"> € Selon "The Guardian", le texte signé par le président Hamid Karzaï, mais pas encore publié, légalise le viol de sa conjointe et interdit aux femmes de sortir sans la permission de leur mari.

<filname="PROF-0,2-3208,1-0,0-2"> € C'est une loi particulièrement rétrograde pour les droits des femmes, "pire qu'à l'époque des talibans", selon la sénatrice afghane Humeira Namati. Elle n'a pas encore été publiée, mais le président Hamid Karzaï l'a signée ces dernières semaines, a indiqué mardi 31 mars The Guardian. Le quotidien britannique explique que ce texte légalise le viol sur sa conjointe et interdit aux femmes de sortir, de travailler ou d'aller chez le médecin sans la permission de leur mari. Selon un document du Fonds de développement de l'ONU en faveur des femmes cité par The Guardian, la loi n'accorde la garde des enfants qu'aux pères et aux grands-pères. Sur le même sujet Hamid Karzaï, le président afghan, et Jan Peter Balekenende, le premier ministre néerlandais, lors de la conférence sur l'Afghanistan organisée mardi 31 mars 2009, à La Haye. Portfolio sonore La communauté internationale à nouveau unie sur l'Afghanistan Le vice-ministre iranien des affaires étrangères, Mohammad Mehdi Akhoundzadeh, mardi 31 mars à La Haye. Compte-rendu L'Iran prêt à coopérer à la reconstruction de l'Afghanistan Les faits La France va envoyer 150 gendarmes en Afghanistan Reportage Comment l'Iran étend son influence en Afghanistan Reportage Visite guidée en "Otanistan" Les faits L'UE se divise sur l'envoi à Kaboul de gendarmes Edition abonnés Thématique : Quelle stratégie en Afghanistan ? Le journal indique que ce texte constitue une concession à la minorité hazara, à quelques mois d'une élection présidentielle qui s'annonce difficile pour Hamid Karzaï. Il cite des élus et représentants d'ONG dénonçant un manque de débats au Parlement sur ce texte contraire à la Constitution afghane, qui garantit des droits égaux pour les femmes. Deux députées mettent l'accent sur le fait que le texte réglemente un domaine jusqu'ici régi par la seule coutume, et qu'il a déjà été amendé (l'âge du mariage pour les filles a été repoussé de 9 ans dans le texte initial à 16 ans).

Les fichiers TXT, sont les plus complets en terme de contenu textuel et d'informations, on décide donc de ne traiter que ce format.

Pour cela on exécute une série d'opérations dont on trouvera le détail ainsi que les principaux raisonnements employés ci-dessous.

1. Sélection des fichiers TXT uniquement et standardisation du contenu. Essentiellement mise sous encodage *Utf8*, et sous format Unix. Ceci car certains fichiers, dont le nombre est difficile à quantifier présentent un encodage *Latin1*

ou encore *Unknown*. Pour ceux que nous arrivons à identifier avec la commande *file -i*, nous modifions l'encodage grâce à l'outil Unix *recode* [5]. Les autres seront laissés au soin de la librairie *Perl*, avec les modules *Encode : :Detect* et *Encode : :HanExtra* [6]
On utilise ici une fonction récursive dont le point d'entrée est le dossier initial. Cette fonction est rappelée si le nouvel argument est un dossier, sinon on lance un traitement particulier si l'argument est un fichier.txt. Tous les autres cas (HTML ou XML) sont ignorés.

2. Uniformisation, grâce à un script perl de chacun des articles, contenus dans les fichiers, sous un format titre/date/article, le tout placé entre balises :
La date, elle, est directement prélevée dans ce type de balise *<AAMMJJ="20150101">*.
On s'appuie sur des balises comme *<filname="SURF-0,2-3208,1-0,0-1">* pour extraire le titre qui apparaît généralement après celles-ci.
L'article, quant à lui, se trouve après le titre selon un schéma variable, mais souvent correspondant à (en couleur le texte que l'on souhaite extraire) :


```
<FILE-date="2009/04/01/19"><article-nb="2009/04/01/19-1"><fi
lnamedate="20090401"><AAMM="200904"><AAMMJJ="2009040
1"><AAMMJJHH="2009040119"><filname="SURF-0,2-3208,1-0,0-1">
 € L'ancien premier ministre Dominique de Villepin o
rganise mercredi en fin d'après-midi une réunion politique s
ur l'OTAN à l'Assemblée nationale.<img width='1' height='1'
 src='http://rss.feedsportal.com/c/205/f/3050/s/3a43153/mf.g
if' border='0'/><br/><br/><a href="http://da.feedsportal.com
/r/35757347385/u/6/f/3050/c/205/s/61092179/a2.htm"><img src=
"http://da.feedsportal.com/r/35757347385/u/6/f/3050/c/205/s/
61092179/a2.img" border="0"/></a><filname="PROF-0,2-3208,1-0
,0-1"> €    <article-nb="2009/04/01/19-2"><filnamedate="2009
0401"><AAMM="200904"><AAMMJJ="20090401"><AAMMJJHH="
2009040119"><filname="SURF-0,2-3208,1-0,0-2"> € Selo
n "The Guardian", le texte signé par le président Hamid Karz
aï, mais pas encore publié, légalise le viol de sa conjointe
 et interdit aux femmes de sortir sans la permission de leur
 mari.<img width='1' height='1' src='http://rss.feedsportal
.com/c/205/f/3050/s/3a42323/mf.gif' border='0'/><br/><br/><a
 href="http://da.feedsportal.com/r/35757345415/u/6/f/3050/c/
205/s/61088547/a2.htm"><img src="http://da.feedsportal.com/r
/35757345415/u/6/f/3050/c/205/s/61088547/a2.img" border="0"/
></a><filname="PROF-0,2-3208,1-0,0-2"> €
C'est une loi particulièrement rétrograde pour les droits de
s femmes,    "pire qu'à l'époque des talibans", selon la sén
```


5. Avec la commande *recode -d latin1..utf8 fichier.txt* que l'on généralise par un appel via une fonction récursive en Perl.

6. voir http://search.cpan.org/~jgmyers/Encode-Detect-1.01/Detect.pm
et http://search.cpan.org/~autrijus/Encode-HanExtra-0.10/lib/Encode/HanExtra.pm. Ces librairies permettent la détection de tous les encodages restants mais certains caractères seront néanmoins mal ré-encodés, lorsque nous rencontrons un exemple de ce type nous le corrigeons (en perl, avec une table de hash *:'猫'=>'è',* puis via une expression du type *$ligne =~s/$motif/$signe{$motif}/g ;*), ceci, au parcours des données, reste malgré tout occasionnel.

atrice afghane Humeira Namati. Elle n'a pas encore été pu bliée, mais le président Hamid Karzaï l'a signée ces dern ières semaines, a indiqué mardi 31 mars The Guardian. Le quotidien britannique explique que ce texte légalise le v iol sur sa conjointe et interdit aux femmes de sortir, de tr availler ou d'aller chez le médecin sans la permission de leur mari. Selon un document du Fonds de développement d e l'ONU en faveur des femmes cité par The Guardian, la lo i n'accorde la garde des enfants qu'aux pères et aux gran ds-pères. Sur le même sujet Hamid Karzaï, le président afghan, et Jan Peter Balekenende, le premier ministre néerlandais, lors de la conférence sur l'Afghanist an organisée mardi 31 mars 2009, à La Haye. Portf olio sonore La communauté internationale à nouveau unie s ur l'Afghanistan Le vice-ministre iranien des affaires étrangères, Mohammad Mehdi Akhoundzadeh, mardi 3 1 mars à La Haye. Compte-rendu L'Iran prêt à coop érer à la reconstruction de l'Afghanistan Les fait s La France va envoyer 150 gendarmes en Afghanistan Reportage Comment l'Iran étend son influence en Afghani stan Reportage Visite guidée en "Otanistan" Les faits L'UE se divise sur l'envoi à Kaboul de gendarm es Edition abonnés Thématique : Quelle stra tégie en Afghanistan ? Le journal indique que ce texte constitue une concession à la minorité hazara, à qu elques mois d'une élection présidentielle qui s'annonce d ifficile pour Hamid Karzaï. Il cite des élus et représentant s d'ONG dénonçant un manque de débats au Parlement sur ce texte contraire à la Constitution afghane, qui garantit des droits égaux pour les femmes. Deux députées mettent l 'accent sur le fait que le texte réglemente un domaine ju squ'ici régi par la seule coutume, et qu'il a déjà été amend é (l'âge du mariage pour les filles a été repoussé de 9 a ns dans le texte initial à 16 ans).
<article-nb="2009/04/01/19-3"><filnamedate="200904
01"><AAMM="200904"><AAMMJJ="20090401"><AAMMJJHH="20
09040119"><filname="SURF-0,2-3208,1-0,0-3"> € L'évol
ution du taux de chômage dans la zone euro et l'ensemble de
l'Union européenne depuis février 2008.<img width='1' heigh
t='1' src='http://rss.feedsportal.com/c/205/f/3050/s/3a42325
/mf.gif' border='0'/>

<a href="http://da.feedsporta
l.com/r/35757345414/u/6/f/3050/c/205/s/61088549/a2.htm"><img
src="http://da.feedsportal.com/r/35757345414/u/6/f/3050/c/2
05/s/61088549/a2.img" border="0"/><filname="PROF-0,2-320
8,1-0,0-3"> €

Ces schémas variables issus de variations textuelles, parfois en raison d'un choix de présentation du journaliste, d'un changement de mise en page par contraintes rédactionnelles, ou encore de l'absence d'un symbole séparateur, peuvent occasionner sur une année, et à plus grande échelle sur 74 mois une

extraction incomplète (on remarque ci-dessus que, par exemple, certains titres n'ont pas d'articles associés). Nous privilégions des expressions régulières généralistes et décidons de quantifier la perte occasionnée dans le tableau volumétrie du corpus. Cette perte se caractérise par un article non extrait quand il ne correspond pas au schéma d'extraction que nous avons établi (non présence et/ou non reconnaissance par nos expressions régulières d'un des trois éléments que sont : date, texte de l'article et titre).

3. Réunion de tous les fichiers en un seul fichier txt par année. Parcours de chaque année et suppression des doublons. Association de chaque article à un id. Et enfin nettoyage :
   On s'appuie sur les titres pour définir les doublons. Ainsi si deux articles partagent le même titre, le second n'est pas extrait. Puisque chaque article est unique, on peut lui associer un identifiant. On profite de ce dernier passage pour effectuer un nettoyage (javascript restant, balise non supprimées, symboles utf8 mal ré-encodés, espace insécable, etc).

Nous verrons dans le chapitre discussion les limites de ce nettoyage, et de cette extraction.

Ces traitements nous permettent de quantifier notre corpus ainsi que les pertes d'articles occasionnées par la non reconnaissance de nos expressions régulières :

Année	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015(2 Mois)**	**Total**
Poids fichiers								
Initial (Go)	1,7	1,3	1,5	1,6	1,9	2,2	287,8(Mo)	10,48
Intermédiaire (avec doublons, sans nettoyage)(Mo)	503	338,1	440,8	422,7	426,9	402	55,3	(Go) 1,26
Final (Mo)	86	66	106	109	94	89	14	(Go) 0,55
Articles								
Nombre d'articles (Réel)	21911	16038	26641	27555	25553	23221	3347	144366
Moyenne (articles/jour)	60,03	43,93	72,98	75,49	70,01	63,62	57,45	63,35
Mots (Millions)	16,8	12,7	20,7	21,4	18,7	17	2,5	109,80
Expressions Régulières								
Perte (%) d'articles non-extraits	2,24	2,36	2,40	4,61	6,52	1,98	1,48	(% moyen) 3,08

TABLE 1.2 – Volumétrie du corpus après traitements

Ces pertes nous apparaissant comme négligeables au regard d'un traitement global du corpus, nous décidons de laisser tel quel le traitement, sans spécialisation supplémentaire de nos expressions régulières.

Nous prévoyons aussi plusieurs sorties de chaque fichier année sous plusieurs formats : TXT (Utf8 et Latin1 voir *cf infra* lexico 3), CSV et XML. Chacun de ces formats étant utilisé dans des contextes différents pour différents tests ou traitements que nous ne détaillerons pas tous ici.

A titre d'exemple on trouvera ici un article de 2009. On pourra y observer notre structure de base : num (numéro d'article), titre, date, article ainsi que le symbole §, utilisé comme délimiteur :

```
§
num : 1
titre : Selon "The Guardian", le texte signé par le président Hamid Ka
rzaï, mais pas encore publié, légalise le viol de sa conjointe et inte
rdit aux femmes de sortir sans la permission de leur mari.
date : 2009.04.01
article : C'est une loi particulièrement rétrograde pour les droits de
```

```
s femmes, "pire qu'à l'époque des talibans", selon la sénatrice afghan
e Humeira Namati. Elle n'a pas encore été publiée, mais le président H
amid Karzaï l'a signée ces dernières semaines, a indiqué mardi 31 mars
 The Guardian. Le quotidien britannique explique que ce texte légalise
 le viol sur sa conjointe et interdit aux femmes de sortir, de travail
ler ou d'aller chez le médecin sans la permission de leur mari. Selon
un document du Fonds de développement de l'ONU en faveur des femmes ci
té par The Guardian, la loi n'accorde la garde des enfants qu'aux père
s et aux grands-pères. Sur le même sujet Hamid Karzaï, le président af
ghan, et Jan Peter Balekenende, le premier ministre néerlandais, lors
de la conférence sur l'Afghanistan organisée mardi 31 mars 2009, à La
Haye. Portfolio sonore La communauté internationale à nouveau unie sur
 l'Afghanistan Le vice-ministre iranien des affaires étrangères, Moham
mad Mehdi Akhoundzadeh, mardi 31 mars à La Haye. Compte-rendu L'Iran p
rêt à coopérer à la reconstruction de l'Afghanistan Les faits La Franc
e va envoyer 150 gendarmes en Afghanistan Reportage Comment l'Iran éte
nd son influence en Afghanistan Reportage Visite guidée en "Otanistan"
 Les faits L'UE se divise sur l'envoi à Kaboul de gendarmes Edition ab
onnés Thématique : Quelle stratégie en Afghanistan ? Le journal indiqu
e que ce texte constitue une concession à la minorité hazara, à quelqu
es mois d'une élection présidentielle qui s'annonce difficile pour Ham
id Karzaï. Il cite des élus et représentants d'ONG dénonçant un manque
 de débats au Parlement sur ce texte contraire à la Constitution afgha
ne, qui garantit des droits égaux pour les femmes. Deux députées mette
nt l'accent sur le fait que le texte réglemente un domaine jusqu'ici r
égi par la seule coutume, et qu'il a déjà été amendé (l'âge du mariage
 pour les filles a été repoussé de 9 ans dans le texte initial à 16 an
s).
\§
```

1.4 Résumé de l'étape et technologies employées

On trouvera ci-dessous le résumé schématique de l'étape, dont les traitements réalisés sont effectués en Perl ou à l'aide d'outils Unix[7].

7. voir section 1.3 p.14.

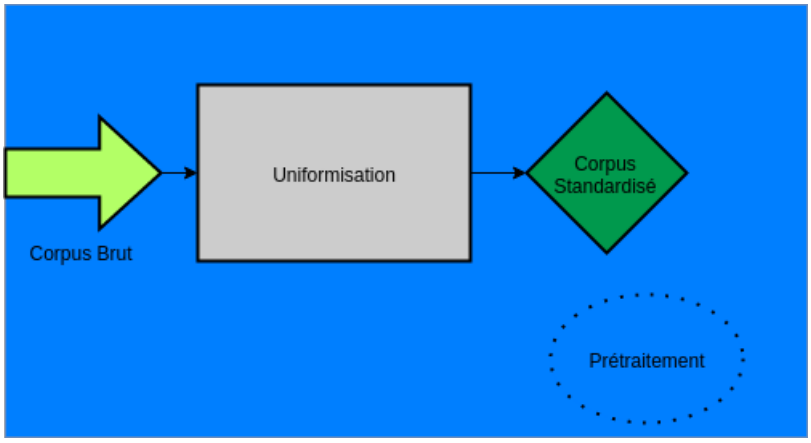


FIGURE 1.2 – Résumé de la première étape des traitements

1.5 Première exploration thématique du corpus

La problématique utilisateur à laquelle nous nous sommes attelés inclus de devoir simuler le regard d'un utilisateur expert. Pour cela une exploration du corpus est nécessaire. Familier de l'utilisation de certains outils textométriques, nous en faisons usage ici afin de définir une liste de vocabulaire permettant d'approcher un thème que pourrait souhaiter étudier notre expert. Au terme de cette étape parallèle nous aurons défini une liste de mots-clefs en rapport avec un thème prégnant de l'actualité.

Cette approche exploratoire est abordée via la textométrie (lexicométrie, ou encore analyse de données textuelles) [8] s'appuyant sur la représentation d'un texte sous la forme d'un cadre et d'une trame pour effectuer des calculs [9]. Pour cela des logiciels comme TXM [10], Lexico3 [11] ou encore le Trameur [12] peuvent remplir ce rôle. Ayant déjà utilisé le Trameur et Lexico3, nous décidons de les employer pour réaliser cette exploration.

Pour cette phase d'exploration test, nous portons notre regard sur les deux premiers mois de 2015, borne temporelle maximale de notre corpus.

Notre problématique est fondée sur la notion de « thème ». Il existe une vaste littérature sur celui-ci en recherche d'information mais nous nous contenterons dans un premier temps de nous limiter à une liste de termes clés afin d'en identifier un. Cette stratégie minimale, qui nous semble suffisante pour commencer à explorer le problème posé, est facilement compréhensible par un utilisateur non expert et pourra éventuellement être complexifiée ultérieurement. Les mois de janvier et février 2015, utilisés pour cette expérience montrent un sujet prégnant dans l'actualité, à savoir le **terrorisme**. Il semble que malheureusement, cette thématique soit porteuse de diverses facettes intéressantes pour notre étude. En effet, outre l'aspect multidimensionnel de cette notion (spatiale, temporelle) et ses évolutions sémantiques, on

8. Voir https://fr.wikipedia.org/wiki/Analyse_de_donn%C3%A9es_textuelles.
9. Pour plus de détail sur ce sujet on consultera le site de l'équipe de recherche ENS de Lyon http://textometrie.ens-lyon.fr/spip.php?rubrique80 ou encore l'ouvrage de [Lebart and Salem, 1994] pour un approfondissement.
10. Pour la présentation du logiciel : http://textometrie.ens-lyon.fr/spip.php?rubrique96.
11. On pourra consulter le manuel d'utilisation [Salem et al., 2003] ou encore le site du projet http://www.tal.univ-paris3.fr/lexico/.
12. http://www.tal.univ-paris3.fr/trameur/

imagine aisément un utilisateur désireux de l'explorer sous divers angles. Nous retenons cette thématique comme thème principal pour cette étude, nous chercherons en conséquence les angles originaux sous lesquels elle peut être représentée. Nous essayons ci-dessous de déterminer quels mots caractérisent ce thème lors des deux mois observés.

Exploration via Lexico3 et le Trameur

On utilise certaines fonctions des deux logiciels que nous ne détaillerons pas ici, cette étape n'étant pas au centre de notre travail, mais ayant simplement pour but de le faciliter en produisant une liste de mots-clefs. De plus le procédé de constitution de cette liste étant incrémental, et étalé dans le temps il serait inutile et fastidieux d'en faire ici le détail. On peut néanmoins citer quelques unes des fonctions utilisées : analyse des spécificités, AFC ou encore les fonctions liées à la cooccurrence [13]. En partant de la notion vague de terrorisme, puis en affinant avec des mots, puis des groupes de mots on constitue peu à peu un lexique thématique relatif au terrorisme.

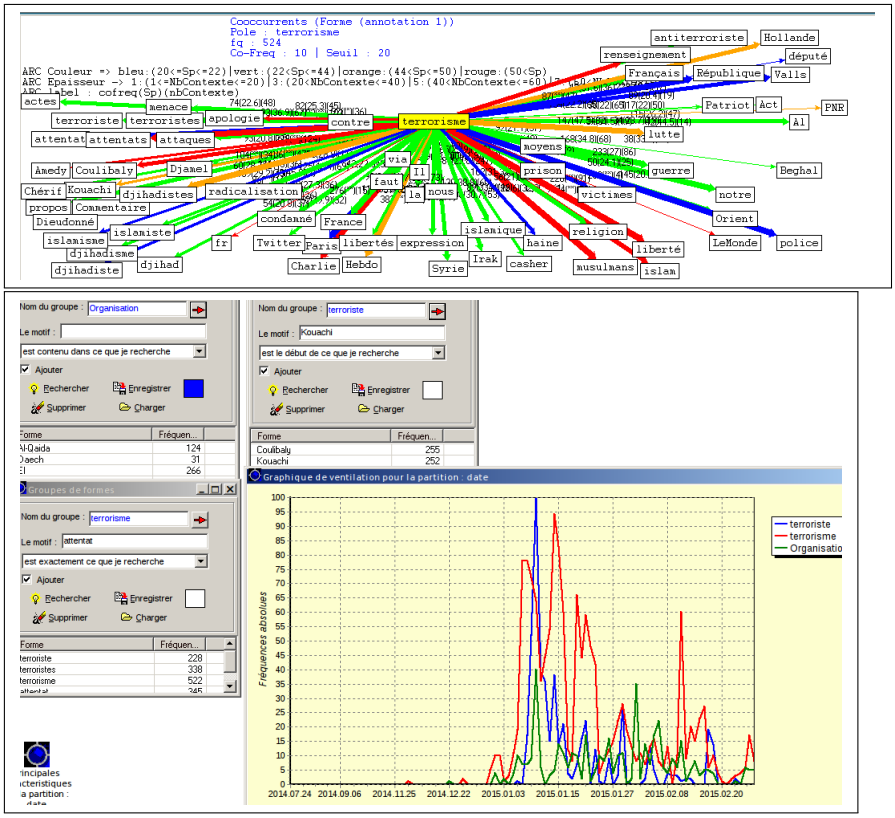


FIGURE 1.3 – Exemple d'utilisation des logiciels : Cooccurrence de terrorisme dans le Trameur et ventilation par groupe de formes selon un axe temps dans Lexico3

Ce travail nous permet de constituer une première liste de 82 termes [14] en relation avec le terrorisme pour les deux premiers mois de l'année 2015. Cette liste nous servira de point de départ pour les explorations thématiques et les simulations de requêtes utilisateurs.

13. On consultera le chapitre 6 de [Lebart and Salem, 1994] pour les spécificités, le manuel de Lexico3 [Salem et al., 2003], ou encore celui du Trameur [Fleury, 2007], pour plus de détails.
14. On trouvera la liste en annexe A p.83.

1.6 Conclusion

Nous avons vu dans ce chapitre les détails du corpus ainsi que les traitements d'uniformisation effectués sur celui-ci. Nous avons aussi survolé la notion de thématique (sur laquelle nous reviendrons) en l'associant à la notion de terrorisme qui sera un fil conducteur de notre scénario utilisateur. Nous avons enfin caractérisé cette thématique en l'associant à une liste de termes issus de l'exploration de notre corpus. Cette phase de préparation terminée, nous pouvons désormais nous consacrer aux traitements liés au filtrage sémantique lors du chapitre suivant.

CHAPITRE 2

MÉTHODES

Sommaire

2.1 Introduction

Ce chapitre traite de trois étapes du traitement global. La première sera l'objet d'un approfondissement de la notion de thème que nous abordions auparavant succinctement. Une fois cette notion explicitée elle nous conduira logiquement à caractériser l'information contenue dans les thèmes en les reliant à une somme d'informations plus restreintes et pertinentes.

Nous poserons comme hypothèse que les entités nommées sont autant de parties de cette information globale recherchée par l'utilisateur. Ainsi, après avoir rappelé les détails de la notion d'entités nommées dans une deuxième partie, nous donnerons deux approches complémentaires pour leur identification. Ces deux approches, l'une basée sur des règles, l'autre par annotation automatique seront ensuite l'objet d'une réflexion sur la complémentarité et le croisement de résultats de façon à optimiser la qualité de l'identification des entités nommées.

Nous verrons enfin dans la troisième partie comment ces entités nommées seront ensuite utilisées comme pivots lors de la visualisation finale, croisées avec la notion

de thématique. Pour cela nous décrirons l'architecture mise en place sous forme d'interface web, conduisant l'utilisateur à entrer des requêtes pour obtenir différentes synthèses visuelles.

2.2 Détection de thème

2.2.1 Présentation

Nous estimons qu'une recherche d'information peut être basée sur des groupements thématiques complémentaires. Ainsi à partir du croisement de thèmes « simples » à définir, on peut délimiter un thème plus « complexe ». Nous aborderons donc dans cette section la question de ce croisement thématique issu de deux thèmes, eux-mêmes caractérisés par des mots-clefs. Ce choix d'utilisation de mots-clefs pour définir un thème simple se base sur trois aspects :

— L'état de l'art de la section suivante montre une certaine complexité à établir la notion de thème
— Dans l'outil que l'on souhaite créer, c'est l'expert, qui définit le thème. L'interaction entre l'utilisateur et l'outil se fait à l'aide d'éléments d'interface qui doivent être simples et accessibles. Un champ de formulaire éditable à l'aide de mots-clés remplit cette condition.
— Les connaissances de l'expert doivent être utilisées pour définir les contours du thème

Le thème complexe, lui, résultant du croisement thématique, permettra en conséquence d'aborder une notion sous différents angles et de synthétiser l'information recherchée suivant différents aspects. Nous commencerons par décrire la stratégie retenue pour la détection et le croisement thématique. Nous aborderons ensuite le traitement par mesures de similarité que nous avons utilisé pour quantifier la proximité entre articles. Enfin nous verrons l'aspect du volume de données et les stratégies pour utiliser le traitement par similarité sur ce volume.

2.2.2 Définition de la notion de thème

On peut souligner la difficulté à trouver un consensus sur la définition de la notion de thème. On trouvera chez [Longo and Todirascu, 2010] le même constat de cette difficulté ainsi que l'ébauche d'explicitation suivante : « *Bien que n'ayant pas de définition unanime, le thème d'un document est considéré en général comme le sujet (" de quoi il s'agit dans un document ") d'une narration, d'un texte explicatif ou d'une conversation.* ». La motivation présentée dans cet article est semblable à la notre [1], pourtant la comparaison s'arrête là. Alors que la démarche de Longo et Todirascu s'attelle à la difficile tâche de l'identification automatique du thème, nous mettons à profit la connaissance que l'utilisateur peut avoir des thématiques qu'il souhaite croiser. En effet, charge à l'utilisateur de définir, par une liste de mots-clés, les contours de ce qu'il considère comme caractéristique de sa thématique. On recherche ensuite les articles en rapport avec la thématique par le biais de ces mots-clés et on les associe à leur *id*. On prend soin de laisser la possibilité, lors du traitement, de choisir le nombre de mot(s)-clé(s) nécessaire(s) au fait de considérer chaque article comme *en*

1. Il y est aussi question d'améliorer l'accès à l'information face à des volumes de données de plus en plus important.

accord ou *hors* thème. Par exemple on peut observer une liste thématique relative au terrorisme de 30 mots-clés telle que celle-ci [2] :

mots-clés Terrorisme	
Al-Qaïda	État Islamique
Amedy Coulibaly	frères Kouachi
AQMI	frères Tsarnaïev
attentat	Hayat Boumeddienne
Ben Laden	je suis Charlie
Boko Haram	jihad
califat	kalachnikov
Charlie Hebdo	kamikaze
Cherif Kouachi	Kenji Goto
Daesh	Mohammed Merah
djihad	organisation terroriste
djihadiste	Saïd Kouachi
Djokhar Tsarnaïev	Tamerlan Tsarnaïev
EI	terrorisme
enlèvement	terroriste

TABLE 2.1 – Liste de mots-clés définissant un thème

On pourrait déjà envisager des améliorations :

— Normalisation des termes présents dans la liste (comme *Tsarnaïev*/*Tsarnaev*, *Al-Qaida*/*Al-Qaïda*, *djihad*/*jihad* ou encore *Cherif et Said Kouachi*/*fréres Kouachi* et même *État Islamique*/*EI*),

— Mettre en place un système de pondération pour établir une hiérarchie d'importance entre les termes

Ces améliorations possibles ne seront pourtant pas traitées ici. En effet, la difficulté inhérente au traitement thématique étant déjà conséquente, sans même envisager ces améliorations.

Cette approche par mots-clés, permet de cerner les notions recherchées par l'utilisateur. On trouvera le schéma explicatif de cette section ci-après, reproduit dans la section concernant l'interface web :

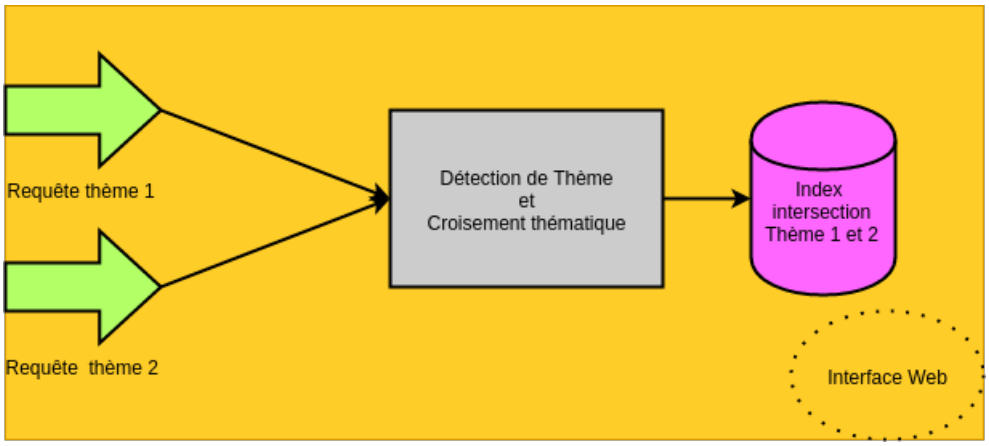


FIGURE 2.1 – Croisement thématique

2. Cette liste est une version simplifiée de la liste présentée en annexe A p.83. En effet la liste de 82 termes étant très orienté vers les attentats de 2015 nous avons limité les variantes de noms propres. De plus il nous apparaît qu'un utilisateur n'utilisera pas d'emblée tous les orthographes d'un mot, ni même autant de mots-clefs sur une première approche.

Les contours de la notion de thème étant flous, on fait l'hypothèse que les articles proches thématiquement (en terme de vocabulaire utilisé) de ceux trouvés seront tout aussi intéressants du point de vue de l'utilisateur. Cette hypothèse permettrait d'agrandir la taille du premier sous-ensemble établi à partir de l'intersection des listes de mots-clés de l'utilisateur. Cet enrichissement se basant sur l'idée simple que des articles partageant le même vocabulaire partageront sûrement la même thématique nous semble pertinent. Il ne reste alors qu'à établir la mesure de calcul adaptée à ce traitement.

2.2.3 Mesures de similarité

Définition

Un certain nombre de calculs sont traditionnellement utilisés en TAL pour évaluer la distance entre deux documents. On peut citer par exemple le TF/IDF ou diverses mesures de similarité comme le coefficient de Dice, l'indice de Jaccard (ou Tanimoto), l'Overlap Coefficient, ou encore la mesure par Cosinus [3].

Utilisés à l'origine en statistique pour comparer la similarité entre échantillons, certains de ces indices ont l'avantage d'être facilement implémentables puisqu'ils reposent sur la notion intersection d'ensembles. Nous avons choisi de mettre en application les mesures suivantes [4] :

— Dice
— Jaccard
— Overlap
— Cosinus

Notre algorithme de traitement repose sur l'implémentation de ces mesures selon les formalisations mathématiques suivantes [5] :

Mesure	Formalisation
Matching coefficient	$\|\mathbf{X} \cap \mathbf{Y}\|$
Dice coefficient	$\frac{\mathbf{2}\|\mathbf{X}\cap\mathbf{Y}\|}{\|\mathbf{X}\|+\|\mathbf{Y}\|}$
Jaccard coefficient	$\frac{\mathbf{2}\|\mathbf{X}\cap\mathbf{Y}\|}{\|\mathbf{X}\cup\mathbf{Y}\|}$
Overlap coefficient	$\frac{\|\mathbf{X}\cap\mathbf{Y}\|}{\mathbf{min}(\|\mathbf{X}\|,\|\mathbf{Y}\|)}$
Cosinus coefficient	$\frac{\|\mathbf{X}\cap\mathbf{Y}\|}{\sqrt{\|\mathbf{X}\|\times\|\mathbf{Y}\|}}$

TABLE 2.2 – Mesures de similarité

On remarque ici par ailleurs que le choix d'implémentation de la mesure de similarité cosinus correspond au coefficient d'Ochiai ceci par souci de simplification : puisque le coefficient d'Ochiai ($K = \frac{n(A\cap B)}{\sqrt{n(A)\times n(B)}}$), où A et B sont des ensembles avec $n(A)$ le nombre d'éléments de A, et si les ensembles sont représentés par des vecteurs de bits, ce coefficient est équivalent à la similarité cosinus

3. On retrouve par exemple des applications récentes de certains de ces calculs de similarité dans la campagne deft2014, actes du 10ème DÉfi Fouille de Textes [Hamon et al., 2014].

4. Encore une fois par manque de temps nous n'aborderons pas la mesure par TF/IDF même si notre intention initiale était de comparer les mesures citées ici et cette dernière. Nous abordons les causes (essentiellement le temps de traitement du volume de données) dans le paragraphe suivant.

5. Issues de [Manning and Schütze, 1999] p.299.

$(cos(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}})$.

Une des contraintes liée à notre corpus, comme nous l'avons déjà mentionné, est la taille de celui-ci. Pour résoudre certains aspects du problème posé, nous avons eu recours à certaines stratégies lors du traitement. Nous présentons l'une d'entre elles ci-dessous.

Application sur volume de données

Pour explorer les thématiques définies par l'utilisateur et observer l'information partagée et contenue dans celles-ci nous nous proposons de croiser deux requêtes avant de les enrichir via les mesures de similarité que nous présentions auparavant. Ainsi, on voit se succéder deux étapes pour répondre à ce besoin :

— une première étape où chacune des deux requêtes répond à une liste de mots-clés définit par l'utilisateur, ainsi qu'à l'exigence de celui-ci d'avoir un certain nombre de mots-clés présents. Les identifiants renvoyés en réponse à ces requêtes correspondent à des textes en rapport avec le thème.

— Suivi d'une seconde où on compare chaque article correspondant aux identifiants renvoyés dans l'étape une à l'intégralité du corpus. On cherche ici le coefficient de similarité le plus élevé. On prend soin de laisser fixer à l'utilisateur le seuil minimal de cette comparaison ainsi que la mesure qu'il souhaite utiliser dans celles proposées.

On imagine bien que la seconde étape, au regard du volume des soixante-quatorze mois d'articles de presse qui compose notre corpus ne peut être calculée en temps réel. On décide d'effectuer un pré-traitement en stockant l'ensemble des résultats dans une base de données. Cette ébauche de solution n'est pas sans soulever d'autres problèmes. En effet au lieu de traiter un nombre variable d'articles renvoyé par chaque requête utilisateur (probablement trop long en temps réel sur l'interface web), on doit désormais calculer le degré de similarité de chaque article avec l'ensemble du corpus. Ce raisonnement revient à un calcul de tous les couples d'un tableau de type N lignes, M colonnes, avec $\mathbf{N} \times \mathbf{M}$ où $\mathbf{N} = \mathbf{M}$. Ce qui revient pour les $\mathbf{N} = 144\,366$ articles composant notre corpus à $\mathbf{N^2} = 20\,841\,541\,956$ combinaisons possibles.

On considère une limitation du nombre d'opérations selon les réflexions suivantes. Si l'on considère une simplification de notre corpus qui ne contiendrait que 10 articles on obtiendrait le tableau suivant :

	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

Dans lequel chaque case vide correspond à un couple d'articles dont on souhaite connaître le degré de similarité et pour lequel un calcul doit donc être effectué soit $\mathbf{N^2 = 100}$.

Cependant, on peut considérer que les couples composés du même identifiant, tels que {1,1}, ne nécessitent aucun calcul puisque leur similarité est maximum. De plus présenter un comparatif de similarité de ce type à l'utilisateur n'aurait pas de sens. On obtient donc la simplification suivante, tel que $(\mathbf{N^2} - \mathbf{N}) = \mathbf{90}$ et pour laquelle une case vide correspond à un calcul à effectuer tandis que le **X** correspond à un calcul n'ayant pas de nécessité :

	1	2	3	4	5	6	7	8	9	10
1	X									
2		X								
3			X							
4				X						
5					X					
6						X				
7							X			
8								X		
9									X	
10										X

De même dans les deux moitiés du tableau considéré, les couples tel que {1,2} sont égaux à {2,1} puisque les mesures de similarité se basent sur l'inclusion. On obtient donc $\frac{(\mathbf{N^2}-\mathbf{N})}{\mathbf{2}} = \mathbf{45}$:

	1	2	3	4	5	6	7	8	9	10
1	X									
2	X	X								
3	X	X	X							
4	X	X	X	X						
5	X	X	X	X	X					
6	X	X	X	X	X	X				
7	X	X	X	X	X	X	X			
8	X	X	X	X	X	X	X	X		
9	X	X	X	X	X	X	X	X	X	
10	X	X	X	X	X	X	X	X	X	X

En appliquant ces règles simplificatrices à notre corpus on obtient pour $\frac{(\mathbf{N^2}-\mathbf{N})}{\mathbf{2}} = \frac{((144\,366 \times 144\,366) - 144\,366)}{\mathbf{2}}$ on obtient malgré tout le résultat de 10 420 698 795.

Ce problème basique en apparence (calcul de la demi matrice sans la diagonale) est intéressant puisqu'il a trait aux stratégies à développer face au volume des données. Or on constate ici que, à raison de 1000 articles par seconde [6], le temps de traitement avoisinerait les 120 jours. Même en parallélisant les processus sur différents cœurs, ou sur différents ordinateurs, nous estimons que le temps de traitement dépassera les délais qui nous sont impartis. Et ce, sans compter le volume des données à stocker. À titre d'exemple on fait un test en comparant les articles de l'année 2009, après 4 jours de traitement (avec 10 traitements lancés en différés) on obtient un fichier de 10Go. Si l'on considère que l'on doit effectué ce travail 28 fois [7], même un traitement différé parait peu envisageable. Pour cette raison nous décidons de procéder à une simplification et d'abandonner l'idée d'un stockage global de tous les

6. Pour une configuration matérielle correspondant à 32Go de RAM pour un processeur de type *Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz*.

7. 2009 avec 2009, puis 2009 avec 2010, ... etc .

résultats de calcul dans une base de données. Au lieu de stocker les résultats de la similarité entre chacun des articles de notre corpus, nous ferons le calcul une fois lors la requête de l'utilisateur.

A ce stade, plutôt que de considérer chaque article comme un élément distinct, nous utiliserons tous les articles définis par l'utilisateur en accord avec le thème qu'il recherche comme un seul article. Pour cela nous établissons une seule fois le dictionnaire des articles concernés comme un seul grand dictionnaire, ce qui revient à comparer ce dictionnaire de mots en rapport avec le thème aux 144 366 articles. Ce parti pris est discutable, en effet si l'un des articles contient un vocabulaire trop proche d'un autre thème, la « balance » peut en quelque sorte être déséquilibrée et le résultat être plus proche du thème le plus représenté lexicalement. Nous discuterons de cet aspect dans le chapitre 4. Compte tenu de ce choix, ne pouvant prévoir les thèmes que choisira l'utilisateur, nous décidons finalement d'appliquer cette méthode en temps réel, dans l'interface web, on trouvera les détails de cette application à la section 2.4.2 p.53.

Cette étape peut être schématisée de la façon suivante, nous ferons rappel de ce schéma dans la section interface pour illustrer nos propos au moment de l'application de la méthode décrite dans cette section :

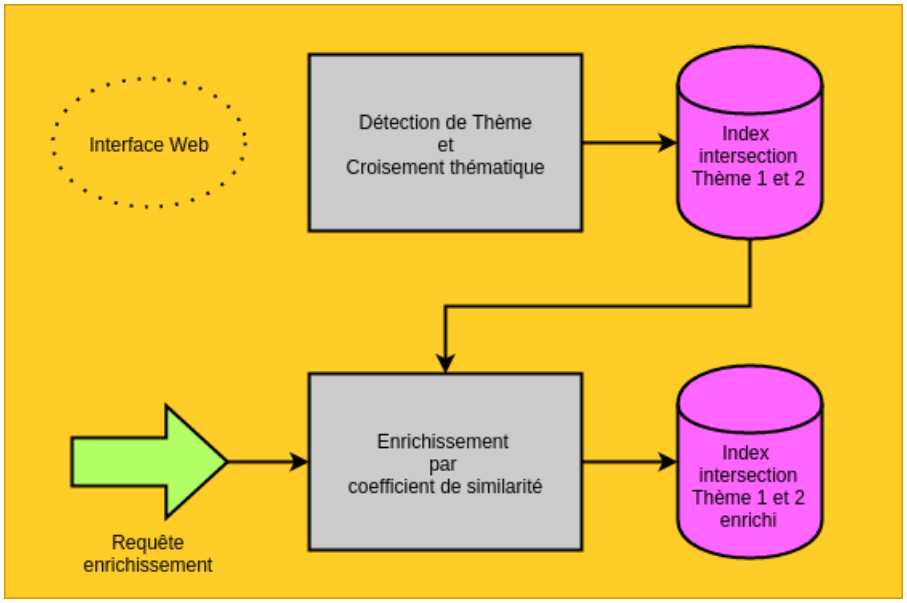


FIGURE 2.2 – Enrichissement thématique

2.3 Entités nommées

2.3.1 Présentation

Une fois un thème détecté, nous avons décidé de le caractériser par différentes parcelles d'informations contenues dans celui-ci. Pour cela nous nous appuyons sur les entités nommées contenues dans ce thème. On trouvera donc dans cette partie, la définition retenue pour la notion d'entités nommées (désormais EN). On décrira ensuite comment l'on procède pour extraire lesdites entités. Pour cela nous utilisons deux approches que nous décrirons successivement :

— La première suit une approche à base de règles, utile pour détecter des structures syntaxiques simples et les EN que celles-ci contiennent dans la phrase, produisant peu de bruit ;
— La seconde utilise des outils d'extraction/annotation automatique d'EN, utiles pour détecter des EN plus complexes mais produisant plus de données bruitées.

Nous détaillerons enfin un système d'indexation associant les EN et les articles dont elles sont issues pour les relier au thème. Nous expliquerons enfin, comment est utilisée cette indexation par le biais de sa transformation en base de données.

2.3.2 Simplification de la notion d'entités nommées

On trouve chez T. Poibeau [Poibeau, 2011], la définition suivante :« *Les entités nommées (EN) désignent les noms de personnes, de lieux, d'organisations, mais aussi les dates ou les unités monétaires.* »[8]. Les différentes campagnes d'annotation et de détection de celles-ci se succédant, le sens associé aux EN s'enrichit et se précise[9].

On peut considérer trois types d'expressions, se subdivisant elles-mêmes en des types plus précis de classification[10] :

— ENAMEX (*entity name expression*)
— TIMEX (*time expression*)
— NUMEX (*numerical expression*)

Dans le cadre de notre scénario utilisateur et pour des raisons de temps ou de difficultés de traitement nous réduirons les éléments recherchés à certains sous-types des deux grandes catégories ENAMEX et TIMEX.

2.3.3 Annotation à base de règles

Avant d'aborder le traitement automatique dans la section suivante, nous allons ici utiliser un outil nous permettant d'effectuer une annotation automatique à base de règles. Ces dernières nous permettent de repérer des EN que nous considérons comme suffisamment simples à étiqueter. On décrit d'abord les fondements théoriques qui nous aident à éditer ces règles puis l'application de celles-ci avec l'outil en question (*DARK*).

Entités ciblées

Beaucoup d'outils permettent aujourd'hui d'extraire les EN, de façon plus ou moins pertinente selon les EN traitées et la complexité de l'extraction. Néanmoins, les plus simples des EN répondant à des schémas qui sont facilement exploitables, pour limiter le bruit et produire un étiquetage rapide avec une certitude de résultat,

8. p.58 pour la citation ou, plus globalement, chapitre 2 p.52-82, pour la question des entités nommées.

9. On trouvera d'ailleurs à ce sujet un descriptif complet des différentes campagnes chez [Nouvel, 2012] p.44. De MUC-6 (1992) à ETAPE (2012), les campagnes sont détaillées suivant les critères suivants :
— Date
— Nom de la campagne
— Langue et modalité de traitement
— Type d'EN retenu
— Métrique d'évaluation

10. [Poibeau, 2011] p.61 .

on décide ici d'utiliser ces schémas grâce à des règles. L'édition de ces règles n'est pas évidente. On commence par un simple repérage lexical à titre de test, avec la notion de pays, un sous-type d'ENAMEX, qui nous donne accès au repérage d'un premier cadre spatial. Ce premier repérage nous permet aussi de faire un test avec certaines librairies graphiques de visualisation et d'évaluer les difficultés [11] séparant les données brutes de la visualisation finale, test que nous détaillerons davantage (voir section 2.4.3 p.59) dans les parties concernées lors des descriptifs sur les différentes visualisations.

On aborde ensuite un repérage temporel plus complexe en s'attaquant à certaines TIMEX. Quand on parle de temporalité, dans la littérature, on voit s'opposer les « *ontologies temporelles* » (points et/ou intervalle) face aux « *ontologies événementielles* »(primauté de l'événement) [12]. Dans les deux cas, apparaît une représentation du temps bornée par des intervalles. Les bornes de ces intervalles peuvent être considérées comme des points précis et nous nous attacherons ici à identifier ces points dans le texte, comme par exemple la mention d'une date telle que « Samedi 7 Février ».

Cependant les règles établies ne feront pas appel ici à la syntaxe, dans ce sens qu'on ne cherche pas à établir un lien entre les TIMEX extraites avec un prédicat ou un événement, comme c'est le cas habituellement dans les ontologies que nous dépeignions, on néglige donc complètement l'aspect verbal. Ce lien pourrait être rétabli ensuite, mais on ne le vise pas explicitement au moment de l'extraction de l'entité. On ne visera pas plus à relier les bornes de ces intervalles, en raison de la difficulté représentée par cette tâche [13]. L'édition des règles est grandement simplifiée par l'usage des travaux de A. Bittar [14] mais surtout par la consultation du guide d'annotation du projet Quaero [Rosset et al., 2011]. Ces travaux, librement adaptés et simplifiés selon nos besoins, seront utilisés dans l'édition des règles que nous expliciterons ci-dessous. De l'application de ces règles apparaîtront des TIMEX que nous appellerons désormais « TIMEX_simple », ceci car notre annotation ne rend pas compte de tous les aspects propres aux TIMEX de façon globale.

DARK

Pour annoter les EN de type TIMEX_simple ou les ENAMEX_pays, nous utilisons l'outil DARK – Data Annotation using Rules and Knowledge (T. Lavergne, LIMSI-CNRS) [15]. Développé en LUA, l'outil permet d'éditer ses propres règles d'annotation. Dans notre usage, nous faisons appel à DARK via un script perl pour chacun de nos articles, stockés sous forme de tableau. Un fichier .lua nous permet d'éditer les règles, pour chaque type d'entités.

Nous privilégions une approche séparative plutôt que globalisante : on a ainsi un fichier pour la configuration ENAMEX_pays et un pour la configuration TIMEX_simple. La facilité d'annotation nous pousse à étendre à un dernier fichier de

11. Difficultés liées essentiellement aux variations d'unités lexicales selon plusieurs langues et formats (voir la note 18).

12. voir [Schwer and Tovena, 2009] pour une explication de ces oppositions.

13. Les intervalles peuvent en effet être présents mais : de formes différentes, implicites, ou même confondus. On imagine bien la variété des cas et, par extension, la variété des traitements qu'il faudrait prendre en compte pour établir les liens entre le début et la fin d'un intervalle.

14. On pourra citer par exemple la méthodologie d'édition du French Time Bank présentée dans [Bittar et al., 2011] de façon synthétique, ou encore le détail du travail réalisé dans [Bittar, 2010].

15. Téléchargeable en version beta sur https://perso.limsi.fr/lavergne/ (dernière consultation le 07/09/15).

configuration TIMEX_metaDate correspondant aux dates d'édition des articles [16].

ENAMEX_Pays
On cherche à obtenir une annotation par article pour chaque pays [17] de type :

```
<PAYS>France</PAYS>
```

Pour cela on associe une liste de pays [18] et une règle dans le fichier de configuration de la forme :

```
main:lexicon("&PAYS",
"~/Bureau/Memoire2015/Scripts/Dark/listePays.txt")
```

Dont la traduction est : annotation entre balises <PAYS></PAYS>de tous les termes présents dans le fichier listePays.txt.

Ce qui donnera par exemple comme résultats [19] :

```
Avec 1,1 million d'hectares cultivés en bio , la <PAYS>France</PAYS>
a dépassé l'<PAYS>Allemagne</PAYS> et affiche la troisième surface
bio d'Europe.
```

TIMEX_simple
On cherche ici à obtenir une annotation par article pour chaque expression temporelle correspondante aux règles suivantes :

```
-- ### Régles Date, entités TIMEX
-- 1) Variables de jour/mois
mois="(01|02|03|04|05|06|07|08|09|10|11|12)"
jour="(1|01|2|02|3|03|4|04|5|05|6|06|7|07|8|08|9|09|10|11|12|13|14|
15|16|17|18|19|20|21|22|23|24|25|26|27|28|29|30|31)"
-- 2) Règles sans dépendance
-- 2.a) Jour de la semaine : lexique
main:lexicon("&WEEK",
"~/Bureau/Memoire2015/Scripts/Dark/configLua/Listes/listeJour.txt")
-- 2.b) Mois de l'année : lexique
```

16. Stockées par nos soins dans un format aaaa/mm/jj nous permettant de séparer les TIMEX_simple des dates métadonnées.

17. Note : On décide de supprimer le nom *Dominic* et le nom *Maurice* de cette liste de pays en raison du grand nombres homonymes dans le corpus (ex : le prénom Dominic) et par rapport à la faible proportion de chance de trouver l'emploi de Dominic/Maurice dans le sens de pays.

18. La liste utilisée est présente sur le site : http://sql.sh/514-liste-pays-csv-xml (dernière vérification le 07/09/15). Le choix de cette liste tient en son exhaustivité (à notre connaissance l'intégralité des pays reconnus par la communauté internationale y est présent) et, de plus à sa présentation selon le modèle suivant :
- id
- code
- alpha 2(ISO 3166-1)
- alpha 3
- nom fr
- nom en gb

En effet, les librairies graphiques utilisées pour faire de la visualisation font souvent appel à l'usage de termes anglais ou de code en alpha 2 ou 3. Ainsi, cette liste a l'avantage de nous permettre à partir du mot français extrait de faire une conversion facile dans le format souhaitée.

19. Tiré de l'article du 2015.02.19.

```
main:lexicon("&MONTH",
"~/Bureau/Memoire2015/Scripts/Dark/configLua/Listes/listeMois.txt")
-- 2.c) Année : patron
main:pattern('[&YEAR (/19%d%d/|/20%d%d/)]')
-- 3) Règles en dépendance
-- 3.a) année => mois
main:pattern('[&MONTH '..mois..'].&YEAR')
-- 3.b) mois => nbr jour du mois
main:pattern("[&DAY 1er] &MONTH")
main:pattern('[&DAY '..jour..'].&MONTH')
main:pattern('[&DAY '..jour..'] &MONTH')
main:pattern('&WEEK [&DAY '..jour..']')
-- 4) Règle finale de l'entité Timex

main:pattern("[&TIMEX (&WEEK &DAY &MONTH &YEAR | &WEEK &DAY &MONTH |
&DAY &MONTH &YEAR | &DAY.&MONTH.&YEAR | &DAY &MONTH |
&MONTH &YEAR | &WEEK &DAY |&WEEK | &MONTH | &YEAR )]")
```

Nous commentons ici les règles 1 à 4 nous permettant le repérage des TIMEX_simple :

1. Variables de jour/mois (utile pour DAY/MONTH) : on initialise deux variables conteneurs de toutes les formes numérales que peuvent prendre les mois ou les jours rencontrés dans le texte.
2. Règles sans dépendance : on considère ici que les éléments ne dépendent pas d'autres éléments pour être caractérisés
   a) Jour de la semaine (WEEK) : un lexique simple des jours de la semaine (lundi, mardi, etc) mais prenant en compte les spécificités de type minuscule/majuscules ainsi que la possibilité que le jour soit suivi de ponctuation.
   b) Mois de l'année (MONTH) : Similaire au lexique précédent mais cette fois avec les mois de l'année.
   c) Année (YEAR) : On fait l'hypothèse ici que l'élément composé des séquences textes "19" ou "20" suivis de deux *digits* sera une année. On assume le risque d'erreur que cela peut provoquer.
3. Règles en dépendance : On considère que les éléments issus des règles suivantes sont des ensembles des règles simples spécifiées auparavant.
   a) La détection d'une année précédée d'un point et d'un chiffre contenu dans la variable mois entraîne l'étiquetage de ce chiffre en mois ("."+"YEAR"⇒"MONTH").
   b) Quatre cas similaires et bâtis sur le même modèle de détection du mois qui entraîne l'étiquetage du jour soit si :
   le mois est précédé d'un espace lui-même précédé de la mention 1er ("1er" + "_ "+"MONTH"⇒"DAY"),
   le mois est précédé d'un point et précédé lui-même d'un élément de la variable jour("variable_jour"+"."+"MONTH"⇒"DAY"),
   le cas 3 est identique au cas deux si l'on remplace le point par une espace("variable_jour"+"_"+"MONTH"⇒"DAY").
   Le cas 4 est semblable : est un DAY un chiffre de la variable jour mais précédé d'un élément étiqueté WEEK.

4. Règle finale de l'entité Timex : une règle complexe qui prend appui sur l'ensemble des détections précédentes, est étiqueté TIMEX avec détection des éléments dans l'ordre, de façon exclusive l'un des éléments suivants :
   — WEEK DAY MONTH YEAR
   — WEEK DAY MONTH
   — DAY MONTH YEAR
   — DAY.MONTH.YEAR
   — DAY MONTH
   — MONTH YEAR
   — WEEK DAY
   — WEEK
   — MONTH
   — YEAR

On peut observer par exemple le résultat [20] :

```
Dans l'usine de Subaru à Ota, province de Gunma, au Japon. La producti
onindustrielle au Japon s'est effritée de 0,7 % en <TIMEX><MONTH>août<
/MONTH></TIMEX> sur un mois. Mais les autorités n'y ont vu, <TIMEX><WE
EK>lundi</WEEK> <DAY>30</DAY> <MONTH>septembre,</MONTH></TIMEX> qu'un
accident deparcours, à la veille d'un discours attendu du premier mini
stre Shinzo Abe,qui devrait annoncer une hausse de la taxe sur la cons
ommation. En <TIMEX><MONTH>août,</MONTH></TIMEX> la production a légèr
ement diminué à cause d'une fabrication moins intense de voitures et d
e semi-conducteurs, selon le ministère de l'industrie. Qui a toutefois
 estimé qu'elle continuait de montrer "des signes de reprise à un ryth
me modéré". En outre, les professionnels s'attendent à un fort rebond
dès le mois de <TIMEX><MONTH>septembre,</MONTH></TIMEX> avec une produ
ction attendue en hausse de 5,2 % d'un mois sur l'autre.
```

Meta-données
En plus des TIMEX_simple on cherche aussi à extraire les dates se rapportant à l'écriture de l'article. On a pris soin pendant l'étape d'uniformisation du corpus de les extraire sous le format *aaaa.mm.jj*. On emploit donc la règle suivante :

```
main:pattern("date ':' [&DATEMETA /%d%d%d%d.%d%d.%d%d/]")
```

Qui traduit l'étiquetage selon l'aspect suivant : est une DATEMETA, un élément composé de quatre *digits* suivis d'un point , de deux *digits* suivis d'un point et de deux derniers *digits*, le tout précédé de la séquence *"date : "* qui correspond à la manière dont nous avons construit notre représentation des données lors de l'étape d'uniformisation.

Ce qui donne comme résultat pour chaque article :

```
titre : Un site publie un échange de lettres entre Marlon Brando et Te
nnessee Williams pour la préparation de la pièce "La Descente d'Orphée
". Une plongée passionnante dans les relations entre l'acteur et le dr
amaturge.date : <DATEMETA>2010.04.01</DATEMETA>article : (EMBED) L'ima
ge de l'acteur Marlon Brando restera à jamais associée à celle de Stan
ley Kowalski, le héros viril d'Un tramway nommé désir (1947)
```

20. article du 2013.09.30

Résultats
On a ainsi étiqueté une collection de repères temporels, spatiaux et de méta-données utiles à la visualisation. On peut quantifier l'ensemble des résultats :

Année	**ENAMEX_pays**	**TIMEX_simple**	**Nbr d'articles**	**Poids fichier** (Mo)	**Perte** (%)
2009	64354	170856	21911	86.9	0.000
2010	50575	124256	16038	68.4	0.019
2011	69261	214138	26641	110.4	0.034
2012	69857	240986	27555	113.9	0.036
2013	66290	224248	25553	101.2	0.063
2014	59021	207133	23221	92.5	0.159
2015 (Jan-Fev)	9427	31839	3447	14.1	0.145
Total	388785	1213456	144366	587.4	(moyenne) 0.065
Total EN			1746607		

TABLE 2.3 – EN extraites via DARK

Remarque : On se rend compte assez rapidement d'un bug « *error : src/-dark.lua :315 : overlong sequence* », dû à une limite de mémoire fixée à 4096 octets. Cette erreur se produit dès qu'une séquence de texte d'article est trop importante par rapport à cet espace mémoire réservé. En raison du faible pourcentage que représente celle-ci et par manque de temps, on décide de faire report de la quantité mais sans correction [21].

2.3.4 Extraction automatique

Nous le mentionnions plus haut, beaucoup d'outils sont disponibles pour l'étiquetage d'EN en français. On peut mentionner de manière non exhaustive :

— la plate-forme issue du projet collaboratif Gate [22],
— le système d'annotation automatique d'EN de D. Nouvel, mXS [23],
— le multifonctionnel projet européen opeNER [24]
— Ou encore la cascade de transducteur CasEN pour la reconnaissance des entités nommées [25]

La plate-forme Gate s'insérant mal, selon nous, dans la démarche de chaîne de traitement globale, celle-ci étant, de plus, assez complexe en terme de prise en main, nous décidons d'orienter nos recherches uniquement vers mXS et le projet opeNER. Enfin par manque de temps, et, de plus, le travail d'mXS s'inspirant en partie de CasEN (pour la partie orientation connaissances) nous faisons le choix de ne pas expérimenter ce dernier, nous n'en ferons pas davantage mention ici.

mXS

Développé dans le cadre de sa thèse de doctorat par Damien Nouvel, [Nouvel, 2012], mXS permet une annotation automatique des EN, en français, basé sur un modèle appris sur le corpus Etape [26] en s'appuyant sur l'extraction de règles d'annotation (par fouille de patrons en utilisant à la fois des). Le fonctionnement

21. Celle-ci étant par ailleurs corrigible en modifiant la mémoire allouée directement dans le code source.
22. https://gate.ac.uk/projects.html
23. http://damien.nouvels.net/fr/mXS
24. http://www.opener-project.eu/index.html
25. http://tln.li.univ-tours.fr/Tln_CasEN.html
26. Pour plus de détails sur le corpus on pourra consulter [Gravier et al., 2012].

de mXS décrit comme utilisant à la fois des « approches orientées connaissances et orientées données » est présenté par Nouvel dans son architecture globale de la façon suivante :
« *[...] le processus que nous cherchons à mettre en œuvre peut être vu comme un apprentissage automatique [...] et comporte donc deux phases distinctes. La première étape est le paramétrage du modèle à partir de données pour lesquelles les entités nommées sont connues, soit l'extraction des règles d'annotation et l'estimation des paramètres d'un modèle numérique utilisant ces règles, [...]. La seconde étape utilise ces paramètres au sein d'un système pour réaliser une prédiction, dans notre cas une annotation, de données pour lesquelles nous cherchons à reconnaître les entités nommées.* » [27].

Cette architecture est résumée dans les figures suivantes [28] :

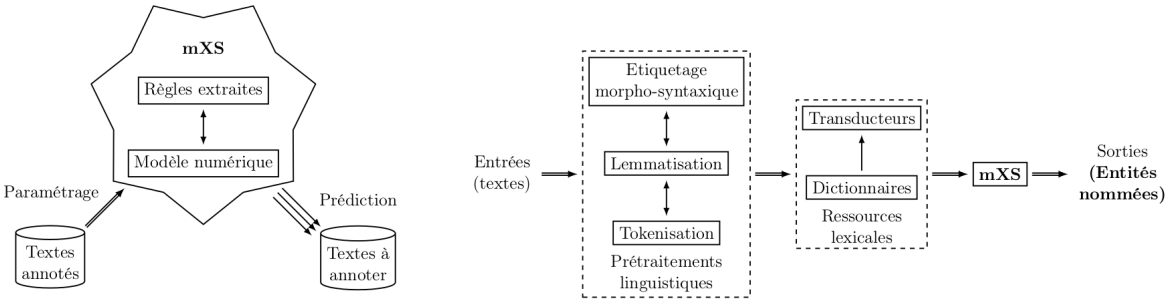


FIGURE 2.3 – Paramétrage et prédiction de mXS et Architecture des traitements d'mXS

Le corpus Etape, quant à lui, est décrit par Nouvel de la façon suivante : « *Le corpus Etape comporte des enregistrements d'émissions télévisuelles, dont le signal audio est extrait* [...] *Ce sont également des émissions d'information, des dossiers d'actualités, des discussions et des débats.* » [29]. Jugeant que ce modèle est applicable à notre corpus journalistique, nous nous appliquons à l'employer sur un petit échantillon du corpus [30]. On trouve des résultats significatifs :
« [...]

```
Un huis clos entre le <func.ind> <kind> président </kind> </func.ind>
grec <pers.ind> <name.first> Georges </name.first> <name.last> Papandr
éou </name.last> </pers.ind> , <pers.ind> <name.first> Nicolas </name.
first> <name.last> Sarkozy </name.last> </pers.ind> et <pers.ind> <nam
e.first> Angela </name.first> <name.last> Merkel </name.last> </pers.i
nd> au moment du référendum sur l'aide à la <loc.adm.nat> <name> Grèce
 </name> </loc.adm.nat> .
```

[...] » [31].

On distingue bien l'intérêt que ce type d'annotation peut avoir dans notre optique de visualisation. Deux problèmes se posent néanmoins, qui contrastent les résultats.

27. [Nouvel, 2012] p.90 .
28. [Nouvel, 2012] p.90, p.91 .
29. [Nouvel, 2012] p.111 .
30. On appelle mXS pour chaque article, plutôt que globalement de façon à limiter la taille du contexte. Ce qui se révèle une mauvaise stratégie comme nous pourrons le voir davantage dans les remarques complémentaires, ceci car si l'on peut obtenir rapidement un résultat pour chaque article, le temps de traitement global est augmenté.
31. Une partie de l'article daté 2015.02.01

1. Annotation erronée ou incomplète.
2. Annotation manquante, et parfois sur une même EN annotée par ailleurs.

Exemple :
« [...]

```
L' Etat islamique a revendiqué et montré le meurtre de l' otage japona
is <name.first> Kenji </name.first> </pers.ind> Goto dans une vidéo pu
bliée en ligne . Il s' agit de la deuxième exécution d' un otage japon
ais en une semaine , après la mort de Haruna Yukawa dimanche dernier .
 Le premier ministre japonais s' est indigné contre un " acte de terro
risme ignoble » . Une semaine tout juste après l' annonce de l' exécut
ion du Japonais Haruna Yukawa par l' organisation Etat <pers.ind> <nam
e> islamique ( EI </name> ) , le <qualifier> premier </qualifier> <kin
d> ministre </kind> nippon , Shinzo Abe </pers.ind> , est à nouveau ap
paru devant la presse le visage fermé , dimanche 1er février , pour co
ndamner « un acte de terrorisme ignoble » . Le groupe djihadiste a pub
lié en ligne une vidéo montrant et revendiquant  l' exécution du journ
aliste Kenji Goto , samedi 31 janvier .
```

[...] »[32].

Pour résoudre le cas 1 nous envisageons à ce stade un filtrage en utilisant des connaissances extérieures du monde.
Pour l'aspect 2 , la solution que nous proposons de croiser le résultat de l'extraction de plusieurs outils nous semble pertinente. Les modalités de ce croisement ne seront malheureusement que peu approfondies dans ce mémoire par manque de temps. Spécifiquement le système d'enchâssement donne des perspectives de traitement intéressantes comme ici :
« [...]

```
L' <loc.adm.nat> Italie </loc.adm.nat> « ne peut pas sous-estimer la p
ossibilité d' une attaque par l' Etat islamique » , estime le <func.in
d> <kind> ministre </kind> des <org.adm> <name> affaires étrangères </
name> </org.adm> </func.ind> , qui affirme que le pays est prêt à lutt
er contre cette menace .
```

[...] »[33].

Malheureusement, lorsque nous dépassons le cadre d'un petit volume de données, on se heurte au problème de temps de traitement. En effet, quand on lance le traitement sur un échantillon plus gros (2 mois complets), le programme prend 5 jours pour traiter environ 1500 articles. Le nombre complet d'articles pour les deux mois en question étant de 3447 on peut envisager un traitement en plus de 10 jours. Sachant que le volume total de nos données correspond à 74 mois[34], l'intervalle de temps de traitement nous pousse à nous tourner vers une autre source d'extraction en retenant les solutions proposées plus haut.

32. Une partie d'un autre article daté 2015.02.01
33. 2015.02.14
34. Rappel sur les données : corpus de 2009-2014, auquel s'ajoute janvier et février 2015 soit environ 590,5 Mo. L'expérience est menée sur une configuration matérielle correspondant elle, à 32Go de RAM pour un processeur de type *Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz*.

Remarques complémentaires

Après contact avec l'auteur du projet nous pouvons proposer une solution alternative basée sur les points suivants pour réduire le temps de traitement :

— Un appel global de mXS plutôt que plusieurs appels localisés (mXS chargeant un certain nombre d'éléments en mémoire à son appel)
— L'utilisation d'un script [35] chargé d'améliorer la vitesse du process en ne ciblant que des EN plus basique (Personnes, Lieux, Organisations), en omettant les sous catégories et imbrications.
— Un traitement par exécutions en parallèle

Nous n'avons pas pu néanmoins tester ces solutions au moment de l'écriture de ce mémoire.

NERC

Le projet opeNER est des plus attractif, dans ce sens qu'il constitue un projet à grande échelle, fondé à l'initiative de la commission européenne [36].

Il est constitué de plusieurs modules que l'on peut utiliser suivant un modèle de « *pipeline* » (dans notre cas on pourra consulter la littérature sur NERC-fr notamment [Azpeitia et al., 2014]) :

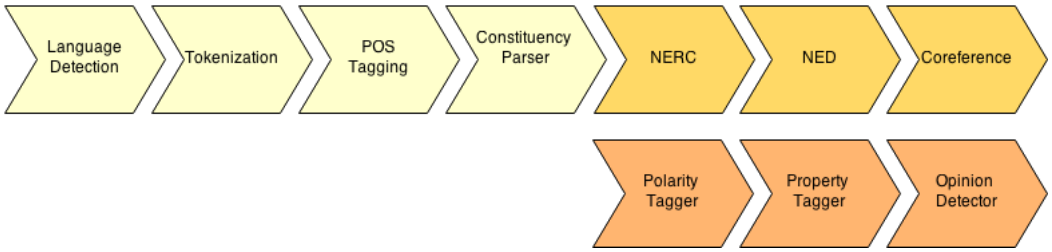


FIGURE 2.4 – Pipeline de présentation du projet opeNER

OpeNER Disposant d'un service web auquel on peut soumettre une requête, et pour laquelle on obtient en retour de celle-ci le texte étiqueté, nous commençons par cette approche (à raison de "1 requête" = "1 article") en adaptant le modèle suivant [37] :

```
cat some_file.txt | curl -F 'input=<-' \
"http://opener.olery.com/language-identifier\
?callbacks\[\]=http://opener.olery.com/tokenizer\
&callbacks\[\]=http://opener.olery.com/pos-tagger\
&callbacks\[\]=http://opener.olery.com/outlet"
```

Néanmoins compte tenu du nombre de requêtes globales que nous devrions soumettre et voyant le faible taux de retour de résultats, (4 résultats pour 10 requêtes en moyenne), nous nous interrogeons sur la démarche. Ces faibles résultats sont-ils dus au nombre de requêtes ? À la taille de celles-ci ? À la structure des données ? Au service web ?

35. Le script *tagEtapeModelPLOP.sh*
36. « *OpeNER is a project funded by the European Commission under the FP7 (7th Framework Program). Its acronym stands for Open Polarity Enhanced Name Entity Recognition. It is a two year duration project which officially started at July 2012, and finishes at July 2014. In OpeNER are collaborating partners from Italy, Holland and Spain.* », tiré de http://www.opener-project.eu/project/
37. voir http://www.opener-project.eu/getting-started/how-to/quick-start.html.

Devant le nombre de ces interrogations sans réponse nous abordons le problème sous un autre angle. En effet opeNER permet une installation en local, ce qui nous permettrait potentiellement une découpe du problème et d'isoler les causes de celui-ci. L'installation d'opeNER simple en théorie, se révèle pourtant ardue en pratique en raison du nombre de technologies en interaction. En effet sous une architecture Unix (Mac et Linux), on retrouve des éléments en :

— Ruby 1.9.3+
— Python 2.7+
— Java 1.7+
— Perl 5+

Dans l'absolu on ne vise que l'utilisation du module qui concerne l'extraction d'EN. On peut, dès lors, faciliter l'opération d'installation en ne s'occupant que de l'élément seul, qui, de plus, existe en java [38]. En consultant la littérature associée, on peut ainsi quantifier les résultats de l'outil pour la version traitant du français. On reproduit, à titre indicatif, les résultats de f-mesure de NERC-fr, consultable chez [Azpeitia et al., 2014] :

Dataset	f1-score without POS	f1-score with POS
Development	88.23	88.34
Test	80.59	80.69

TABLE 2.4 – F-mesure globale de NERC (%)

On garde de plus à l'esprit que les performances de NERC-fr sont variables selon les EN [39] :

Categories	f1-score without POS	f1-score with POS
Location	87.29	87.18
Person	84.41	84.54
Time	85.09	85.45
Date	78.96	79.03
Organization	64.00	64.38
Money	36.44	37.55

TABLE 2.5 – F-mesure par catégorie de NERC (%)

On utilise le même procédé que pour mXS, en faisant appel à NERC on peut ainsi comparer les résultats avec ceux précédemment observés :

— `Un huis clos entre le président grec <START:person> Georges Papandréou , <END> <START:person> Nicolas Sarkozy <END> et <START:person> Angela Merkel <END> au moment du référendum sur l'aide à la <START:location> Grèce.Né <END>`
— `L'Etat islamique a revendiqué et montré le meurtre de l'otage japonais <START:person> Kenji Goto <END> dans une vidéo publiée en ligne. Il s'agit de la deuxième exécution d'un ota`

38. Et donc directement exécutable sans installation préalable autre que java.
39. Reproduction des résultats à nouveau issue de [Azpeitia et al., 2014].

```
ge japonais en une semaine,  après la mort de <START:person>
 Haruna Yukawa <END> <START:date>  dimanche <END> dernier. L
e premier ministre japonais s'est indigné  contre un "acte d
e terrorisme ignoble ». Une semaine tout juste  après l'anno
nce de l'exécution du Japonais <START:person> Haruna Yukawa
 <END> par l'organisation Etat islamique (EI), le premier mi
nistre  nippon, <START:person> Shinzo Abe, <END> est à nouve
au apparu  devant la presse le visage fermé, <START:date> di
manche <END> 1er <START:date> février, <END> pour condamner
« un acte de  terrorisme ignoble ». Le groupe djihadiste a p
ublié en  ligne une vidéo montrant et revendiquant l'exécuti
on du  journaliste <START:person> Kenji Goto, <END> <START:d
ate>  samedi 31 janvier. <END>
```

— `L'Italie « ne peut pas sous-estimer la possibilité d'une att aque par l'Etat islamique », estime le ministre des affaires étrangères, qui affirme que le pays est prêt à lutter contr e cette menace.`

On constate globalement une perte de précision de l'information par rapport à mXS concernant les fonctions (attribut *func* [40]), les fonctions de président ou ministre n'apparaissent plus. On peut aussi souligner certaines non-reconnaissances (cependant ce phénomène se produit pour les deux outils), mais dans des cadres différents. Ces non-reconnaissances montrent la nécessité de croiser l'information issue de plusieurs outils. Les exemples ci-dessus illustre ce propos avec des éléments comme :

```
<START:date> dimanche <END> 1er <START:date> février, <END>
```

Avec le découpage en deux entités de type date ce qui représente une seule et même entité, ou encore le fait qu' *Italie* dans le troisième exemple n'est pas reconnu comme une entité de type *location* [41].

On pourrait aussi avancer une augmentation du nombre de d'entités « *person* [42] » : en observant NERC sur 3447 articles on obtient 30494 entités du sous-type d'ENAMEX *person* contre 5650 *pers.ind* pour mXS à raison de 1532 articles traités. Ces résultats sont à nuancer en raison de trois points :

— En raison du temps de traitement nous avons dû stopper le traitement à 1532 articles pour mXS rendant une comparaison difficile
— Les objectifs des outils, bien que similaires possèdent des différences significatives dans les types d'EN recherchées
— L'augmentation du nombre d'étiquettes *person* chez NERC n'est pas exempt d'étiquettes mal posées : par exemple l'organisation terroriste *Boko Haram* est reconnue comme une *person*. Contre une non-reconnaissance chez mXS.

Dans ces conditions nous souhaiterions utiliser les deux, néanmoins, encore une fois en raison du volume de données nous n'utiliserons que NERC qui traite notre

40. On se reportera à nouveau vers [Rosset et al., 2011] pour le détail des annotations utilisées dans mXS, ou encore vers les chapitres concernant les marqueurs d'annotation présents chez [Nouvel, 2012]. Nous ne ferons pas davantage détail de ces annotations si cela n'est pas nécessaire, puisque, en effet peu seront réellement utilisées dans le contexte qui nous intéresse.
41. type d'entité présent sous NERC pour désigner un emplacement géographique.
42. type d'entité présent sous NERC pour désigner un nom de personne.

corpus en moins d'une heure. On peut ainsi obtenir les entités nommées suivantes [43] :

— **Date**
exemple :« *le nuit de <START :date> samedi 31 janvier <END>* »
— **Location** (localisation spatiale : pays, ville, région administrative, etc)
exemple : « *Les stations de ski des <START :location> Hautes-Pyrénées <END>* , ou encore *de rapatrier ses enfants en <START :location> France <END>* »
— **Money**
exemple : « *plusieurs milliers de particuliers et de PME infectés, et une rançon de <START :money> 300 euros <END> pour chacun à payer.* »
— **Organization**
exemple : « *Mélenchon se rêve en héraut d'un <START :location> Syriza <END> français LE <START :organization> PS <END> "PÉRIMÉ"* »
— **Person**
exemple : « *<START :person> Cécile Duflot <END> a gouverné et en a gardé le goût.* »
— **Time**
exemple : « *chaînes d'information en continu, « <START :time> 20 heures <END> », réseaux sociaux...* »

Le tableau suivant synthétise les résultats extraits par NERC-fr pour ces catégories :

Categories	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015** (Jan-Fev)	**Total**
Date	106363	78071	133412	136438	119165	104821	16812	695082
Location	163523	116226	172667	181355	163821	155923	23630	977145
Money	8504	5855	8005	10816	9143	8859	1304	52486
Organization	45325	33463	55504	66597	55756	49441	7097	313183
Person	225303	155945	250169	262306	215309	198728	30494	1338254
Time	1788	1408	2347	2131	2029	2284	305	12292
Total	550806	390968	622104	659643	565223	520056	79642	3388442

TABLE 2.6 – EN extraites via NERC-fr

On observe au passage que le total des En extraites par la méthode par règles 1746607 additionné à la méthode par extraction automatique 3388442 est de : 5135049 en tout [44].

On doit maintenant organiser les entités extraites de façon à pouvoir les utiliser, en ce basant sur un système d'indexation reliant le thème, les articles et les EN contenues dans ces articles. Pour présenter une visualisation contenant ces EN à l'utilisateur on effectuera des appels aux index des articles détectés comme traitant du thème recherché. De cette façon l'utilisateur obtiendra une synthèse de l'information basée sur les EN dans un cadre thématique.

43. Toujours chez [Azpeitia et al., 2014] concernant les types d'EN étiquetés par NERC :«
— *The geo-social-political tags were divided into three subcategories : gsp.pers, gsp.loc and gsp.org. These subcategories were then placed under person, location and organisation, respectively.*
— *The amount category, and its amount.cur subcategory, were categorised under money.*
— *product named entities were not used in our system.*
— *person, location, organization, time and date types were maintained as is.*
». Par ailleurs l'essentiel des exemples illustratifs fournis ici sont issus d'articles de février 2015.

44. On remarque cependant que certains des neuf types extraits sont sujets à intersection, ces intersections pourraient être utiles lors de la phase de validation des EN.

Par exemple un utilisateur pourra rechercher l'intersection du thème « *Drone* » et « *Terrorisme* » et observer des EN (personnes, organisations, etc) en rapport avec ce thème.

Chaque type d'EN étant d'une nature sémantique différente, on peut imaginer produire différentes synthèses d'informations en fonction du type d'EN. On cherche donc à extraire plusieurs types d'EN.

2.3.5 Indexation

Nous l'avons déjà mentionné, nous considérons les EN comme un atome du sens global. Nous verrons dans la section relative à la visualisation que le choix de traitement par EN procure des avantages pour la structuration car l'on ne traite pas l'information « à plat ». Cette orientation doit néanmoins intervenir très tôt dans la conception de la chaîne de traitement sous peine d'obtenir des informations traitées comme issues de même nature et difficilement représentables. L'alternative, utile dans certains cas comme première approche du corpus, va à l'inverse de ce que nous essayons de produire ici. Nous cherchons, en effet, à atteindre des données ciblées pour l'utilisateur. Cette idée de structuration est à rapprocher d'une conception ontologique, et l'on peut retrouver cette thématique de traitement de l'information en appui sur les EN dans des travaux récents [45].

Cette opération de structuration s'appuie, dans notre cas, sur un système d'indexation. On rapporte en effet les EN extraites à la zone de texte dont elles sont issues. Chacune de ces zones de texte est auparavant associée à un Id, de cette façon les EN d'un certain type et Id sont associés. On peut ainsi stocker, au format CSV, un certain nombre d'informations parcellaires relatives aux EN issues des extractions que nous avons précédemment exposées. Chacun des fichiers CSV représente un type d'extraction associant toutes les EN d'un certain type à tous les articles dont elles sont extraites comme par exemple pour les huit premières lignes du fichier d'ENAMEX_pays de 2015 :

```
1;Israël;France;France,;France.;
3;France.;
5;France,;
6;Cameroun,;Niger,;Tchad;Bénin;
7;Grèce;Espagne,;France;Grèce,;
8;Japon;Jordanie;
9;Suède,;Suède.;
10;Monaco;
```

On voit ici que les articles 1, 3, 5, 6, 7, 8, 10 parlent de pays. Ces pays sont listés par article. Ainsi le premier article de 2015 parle trois fois de la France et une fois d'Israël. On constate par ailleurs que pour les EN extraites des opérations de nettoyage/filtrage/uniformisation pourraient améliorer la qualité des résultats.

Filtrage d'EN

Cette indexation réalisée on peut envisager deux types de filtrage pour valider les informations ou optimiser leur qualité et, ainsi, faciliter la visualisation. Ces filtrages

45. On pourra se reporter à [Omrane et al., 2011] pour un exemple concernant ce type d'approche et notamment l'idée que les ontologies peuvent être peuplées par les EN mais, aussi, à l'inverse, que ces dernières peuvent être une partie intégrante dans la conceptualisation d'ontologies.

qui correspondent à un nettoyage des données peuvent se découper en deux grands types :

— **Filtrage Interne** : basé sur la structure des éléments en eux-mêmes.

— **Filtrage Externe** : basé sur une connaissance extérieure.

Après avoir présentés ces grands types de filtrage nous détaillons les traitements envisagés pour les appliquer aux neuf types d'EN extraits suivant dans les sections suivantes :

1. Extraction à base de règles
   a) Pays
   b) Datemeta
   c) Timex_simple
2. Extraction automatique
   a) Date
   b) Location
   c) Money
   d) Organization
   e) Person
   f) Time

Filtrage Interne

On pouvait observer dans un exemple précédent :

```
1;Israël;France;France,;France.;
```

On constate que l'EN du type *pays* `France` apparaît trois fois et, parfois, accompagnée de ponctuation. Cette constatation est généralisable pour la méthode d'extraction avec règles. Il apparaît le même type d'extraction imparfaite pour la méthode d'extraction automatique [46] :

```
10565 ; Barack Obama, Benyamin Nétanyahou, ; Yasser Arafat ; Mahmoud A
bbas, ; Gilad Shalit ; Abbas ; Ehoud Barak, ; Avigdor Lieberman, ; Ben
oît XVI, ; Barack Obama ; Michel Bôle-Richard Article ;
```

Où l'on constate la présence de doublons, d'espaces, et de caractères de ponctuation [47]. Cependant chacun des ensembles d'EN extraits ne présente pas la même structure d'imperfections. Cette correction est néanmoins envisageable puisqu'elle suit des schémas prédictibles.

Application Filtrage Interne : Datemeta

Les meta-données des dates d'articles ne nécessitent aucun traitement particulier en filtrage interne. En effet toutes de la forme `1;2009.04.01;`, elles sont

46. Exemple d'un article du 2009.06.01 pour des EN du type *person*.
47. Rappel : Ponctuation autre que les points-virgules, séparateurs de nos fichier csv.

directement exploitables.

Application Filtrage Interne : Pays, Location, Organization, Person

Pour ce qui est des EN de type *Pays*, *Location*, *Organization*, *Person*, on peut effectuer un filtrage qui améliore la conformité des EN attendues en supprimant uniquement la ponctuation, les espaces comme chez les EN *Location* :

```
27; L'Amérique ; Irak ; Bagdad. ; Bassora ; Irak. ; Irak, ; Al-Fadhil ; Bagdad
; Vilseck, ; Allemagne ;
```

Application Filtrage Interne : Timex_simple, Date, Money, Time

Par manque de temps, nous n'effectuerons pas de filtrage sur ces EN.

Filtrage Externe

Observons à nouveau l'exemple :

```
10565 ; Barack Obama, Benyamin Nétanyahou, ; Yasser Arafat ; Mahmoud A
bbas, ; Gilad Shalit ; Abbas ; Ehoud Barak, ; Avigdor Lieberman, ; Ben
oît XVI, ; Barack Obama ; Michel Bôle-Richard Article ;
```

Nous avons introduit ici un élément significatif et digne d'intérêt pour nos traitements d'erreurs : `Michel Bôle-Richard Article`, ou le mot Article est considéré comme appartenant à l'EN. On peut observer d'autres types d'erreurs dans l'exemple suivant extrait du type *person* de l'index 2015 :

```
12; Boko Haram ; Boko Haram, Abubakar Shekau ; Crédits ; Boko Haram,
; Jeudi ; Ban Ki-moon ;
```

Il nous permet de souligner les EN extraites `Crédits`, `Jeudi`, et `Boko Haram` qui représentent autant d'erreurs à corriger.

Pour pallier ces problèmes d'extraction on propose d'utiliser des ressources en ligne de type web sémantique. On obtiendrait par exemple à une interrogation sur *Boko Haram* via la base de données dbpedia [48] une réponse de type :

```
<http://fr.dbpedia.org/resource/Boko_Haram>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2002/07/owl#Thing>
```

Ceci nous permettrait de relier l'entité au type *Thing* et non au type *Person* initialement extrait. Cette opération de « retypage » est particulièrement intéressante pour des EN comme :

```
56; Anne-Marie Vaillé, ; Amedy Coulibaly ; Djamel Beghal ; Najat
Vallaud-Belkacem, ; Régis Debray ;
```

Puisque l'on obtient en réponse [49] :

48. Réalisé en local via la commande : `cat instance-types_fr.nt | grep "Boko_Haram"`.
49. Sur des requêtes de même type que dans la note 48. On interroge d'abord le fichier *fr* d'instances de Dbpédia, en cas de non réponse, comme pour Djamel Beghal, on interroge le fichier *en*.

Sujet	Prédicat	Objet
Requête pour : **Anne-Marie Vaillé**		
Acune réponse	Aucune réponse	Aucune réponse
Requête pour : **Amedy Coulibaly**		
<http ://fr.dbpedia.org/resource/Amedy_Coulibaly>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://dbpedia.org/ontology/Criminal>
Requête pour : **Djamel Beghal**		
<http ://dbpedia.org/resource/Djamel_Beghal>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://dbpedia.org/ontology/Person>
<http ://dbpedia.org/resource/Djamel_Beghal>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://wikidata.dbpedia.org/resource/Q5>
<http ://dbpedia.org/resource/Djamel_Beghal>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://xmlns.com/foaf/0.1/Person>
<http ://dbpedia.org/resource/Djamel_Beghal>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://www.w3.org/2002/07/owl#Thing>
<http ://dbpedia.org/resource/Djamel_Beghal>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://schema.org/Person>
<http ://dbpedia.org/resource/Djamel_Beghal>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://wikidata.dbpedia.org/resource/Q215627>
<http ://dbpedia.org/resource/Djamel_Beghal>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://www.ontologydesignpatterns.org/ont/dul/DUL.owl#NaturalPerson>
<http ://dbpedia.org/resource/Djamel_Beghal>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://dbpedia.org/ontology/Agent>
<http ://dbpedia.org/resource/Djamel_Beghal>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://www.ontologydesignpatterns.org/ont/dul/DUL.owl#Agent>
Requête pour : **Najat Vallaud-Belkacem**		
<http ://fr.dbpedia.org/resource/Najat_Vallaud-Belkacem>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://dbpedia.org/ontology/Politician>
<http ://fr.dbpedia.org/resource/Najat_Vallaud-Belkacem__1>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://dbpedia.org/ontology/PersonFunction>
<http ://fr.dbpedia.org/resource/Najat_Vallaud-Belkacem__2>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://dbpedia.org/ontology/PersonFunction>
Requête pour : **Régis Debray**		
<http ://fr.dbpedia.org/resource/Rû0E9gis_Debray>	<http ://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http ://dbpedia.org/ontology/Person>

TABLE 2.7 – Réponses aux requêtes via Dbpedia

Si la requête peut rester sans réponse, elle peut aussi être un facteur important de caractérisation des EN. Dans un premier temps, il faudrait trouver des correspondances entre les instances Dbpedia et nos EN extraites pour vérifier la validité de ces dernières. On éviterait de présenter des entités non valides à l'utilisateur comme les `Crédits` et `Jeudi` que nous présentions auparavant. Il est aussi envisageable d'effectuer le « retypage » que nous avancions pour des entités comme `Boko Haram`, qui passerait alors d'un type *person* à un type plus proche de la réalité lors de la représentation visuelle. L'association d'EN comme *Amedy Coulibaly* à *Criminal*, ou *Najat Vallaud-Belkacem* à *Politician* est aussi souhaitable et permettrait d'accéder à des informations supplémentaires (comme des pages internet relative à nos EN) tout en permettant de lever l'ambiguïté sémantique relative à l'EN (problème d'homonymie). Des travaux dans ce sens apparaissent, notamment on trouve chez [Ruiz and Poibeau, 2015] l'idée de combinaison d'outils et de vote entre ces outils pour optimiser la qualité des EN et désambiguïsé. Cependant nous n'avons pu atteindre le stade ou nous aurions pu faire des tests basés sur cette idée. Néanmoins dans notre cas, compte tenu du nombre élevé d'EN extraites, il n'apparaît pas judicieux d'effectuer ce filtrage à cette étape (cela reviendrait en effet à vérifier la validité de toutes les entités de notre corpus. De plus la relative complexité des bases ontologiques de Dbpedia rend l'accès à l'information difficile. Nous décidons de reporter ce filtrage à l'étape de visualisation, et de travailler uniquement quand le focus thématique de l'utilisateur à été effectué. (*cf infra* Lors de la partie visualisation)

L'ensemble de l'étape d'extraction d'EN pour les relier aux index des articles dont ils sont issus, à nouveau réalisée en Perl peut être schématisée comme suit :

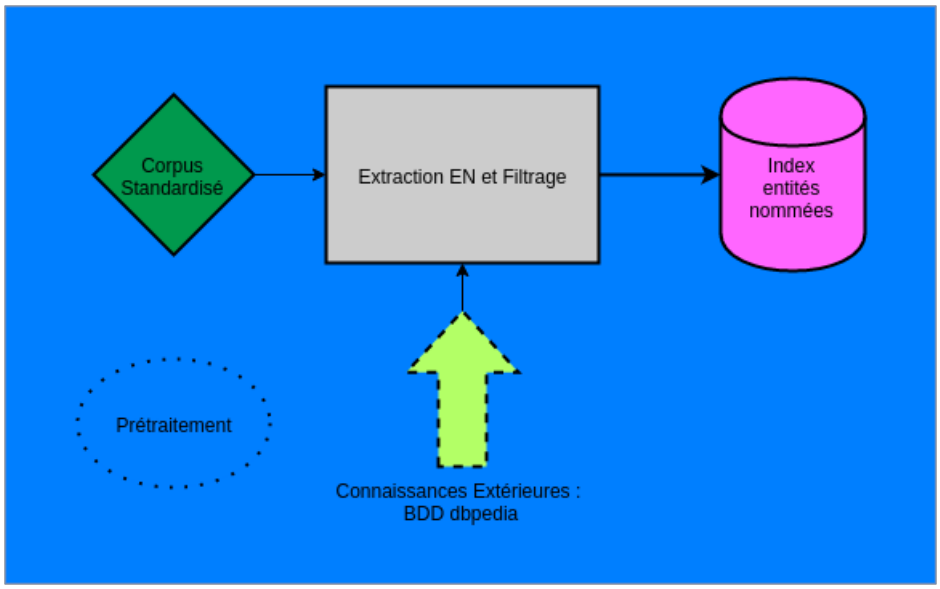


FIGURE 2.5 – Résumé de l'étape Extraction d'EN et Indexation

L'entrée connaissances extérieures apparaît ici en pointillée car nous n'avons pu finaliser l'idée du filtrage décrit plus avant. En interne l'optimisation de la qualité des données en croisant les outils n'a pas pu non plus être réalisée à ce jour.

Base de données

L'appel répété aux index pour diverszq visualisations dans notre interface web entraîne un besoin d'accès rapide avec gestion des accès concurrents, on se tourne naturellement vers les bases de données. La transformation des index en base de données s'effectue via une importation au format *SQlite*.

Le choix de cette bibliothèque ne doit rien au hasard. En effet, facilement intégrable à une application puisque l'accès à une base de données *SQLite* se fait par l'ouverture d'un fichier correspondant en s'affranchissant du paradigme client-server de la majorité des bases de données [50]. Cette décision nous donne une certaine flexibilité sur les différentes mises à jour et autres corrections des données que nous allons avoir à effectuer avant de devoir déployer l'application web. De plus *SQLite* est léger, simple d'utilisation et se justifie particulièrement lorsque il y a peu d'accès aux données en écriture. Ceci est notre cas puisque une fois les index incorporés sous forme de table, nous ne ferons que des consultations sur les données et aucune écriture.

Concernant la structure de la base de données (désormais *BDD*), l'examen de nos fichiers index *CSV* nous montre différents points à évaluer lors de sa création :

— Chaque article représentera une ligne divisée en champs
— On ne connaît pas *a priori* le nombre de ces champs
— On trouve des articles ne possédant aucune extraction de certains type d'EN

Pour éviter de multiples champs *NULL* dans notre base de données et aussi pour des facilités lors du traitement de l'interface web on décide de produire une base de données avec x tables tel que :

— Chaque table correspond à un type d'EN
— Chaque table contient deux champs, *id* et *en*

50. https://fr.wikipedia.org/wiki/SQLite

— les champs *id* représentent les identifiants d'articles
— les champs *en* représentent les EN extraites dans l'article correspondant, sous forme de chaîne de caractères, chaque EN étant séparé de la suivante par un « ; ». On redécoupera la chaîne de caractères au moment de l'affichage en visualisation

Ces points sont résumés dans le schéma suivant :

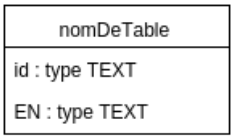


FIGURE 2.6 – Schéma général d'une table de la base de données

L'importation s'effectue de manière très simple en suivant le code suivant :

```
1 create table nomDeTable(id TEXT,en TEXT);
2 .mode csv
3 .separator ";"
4 .import ./fichier.csv nomDeTable
```

On peut, bien sûr, interroger la BDD en SQL une fois celle-ci créée :

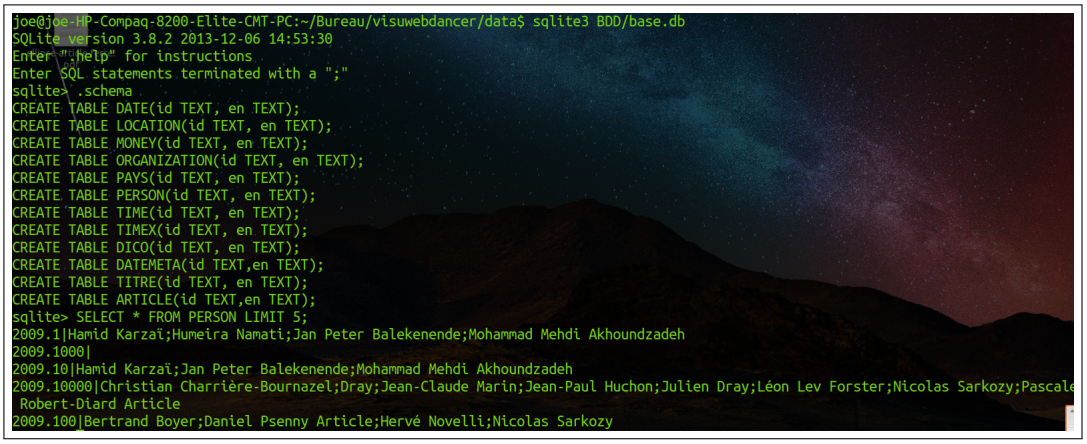


FIGURE 2.7 – Interrogation de la base de données SQL

Enfin cette opération de transformation peut être schématisée simplement de la façon suivante :

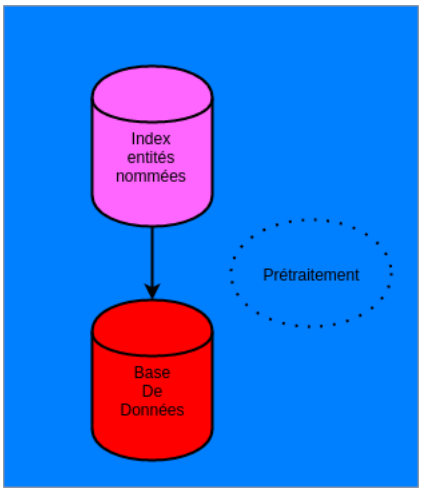


FIGURE 2.8 – Résumé de l'étape passage de l'index d'EN à la BDD SQlite

2.4 Architecture Interface Web

2.4.1 Présentation

Une fois les différents prétraitements effectués, nous pouvons préparer l'interface web qui permettra à l'utilisateur d'accéder à des représentations visuelles des données, affichées en fonction des requêtes thématiques qu'il aura fourni de façon dynamique. Chaque étape de la construction, y compris la solution retenue est présentée ci-dessous. La tâche de cette interface se décline en trois points :

— Permettre à l'utilisateur de rechercher des informations thématiques et de croiser ces thématiques en utilisant deux listes de mots-clés (une liste par thème), la réponse étant une liste d'articles de la forme :

A) Thème 1

B) Thème 2

C) Intersection Thème 1 et Thème 2

— Permettre un enrichissement de cette intersection de thème [51]

— la possibilité à l'utilisateur une fois les thèmes établis, de consulter des informations sur ces thèmes grâce à différentes visualisations ou encore grâce à un retour au texte.

La plupart de nos scripts de test (par exemple concernant la section 2.2.3 pour l'enrichissement) sont en langage Perl. C'est donc naturellement que nous décidons d'implémenter l'interface dans le même langage.

2.4.2 Interface

Nous utiliserons *Dancer* un framework léger de développement d'applications web. *Dancer* est écrit en Perl [52]. Basé sur le paradigme *Modèle-Vue-Controleur* (ou

51. On laisse ici le choix à l'utilisateur de l'utiliser, cependant il est clair qu'un enrichissement sur un grand nombre d'articles ne paraît pas judicieux.

52. Les références ne manquent pas, mais on consultera par exemple :
Le site de Dancer http://perldancer.org/
Sur CPAN http://search.cpan.org/ xsawyerx/Dancer2-0.163000/lib/Dancer2.pm

MVC [53]), il facilite le développement en se basant sur un moteur de templates. On développe notre application basé sur *Dancer2*. Pour plus de détails sur l'implémentation de l'application web on consultera l'annexe F p.95.

L'accueil de l'interface se présente de la façon suivante :

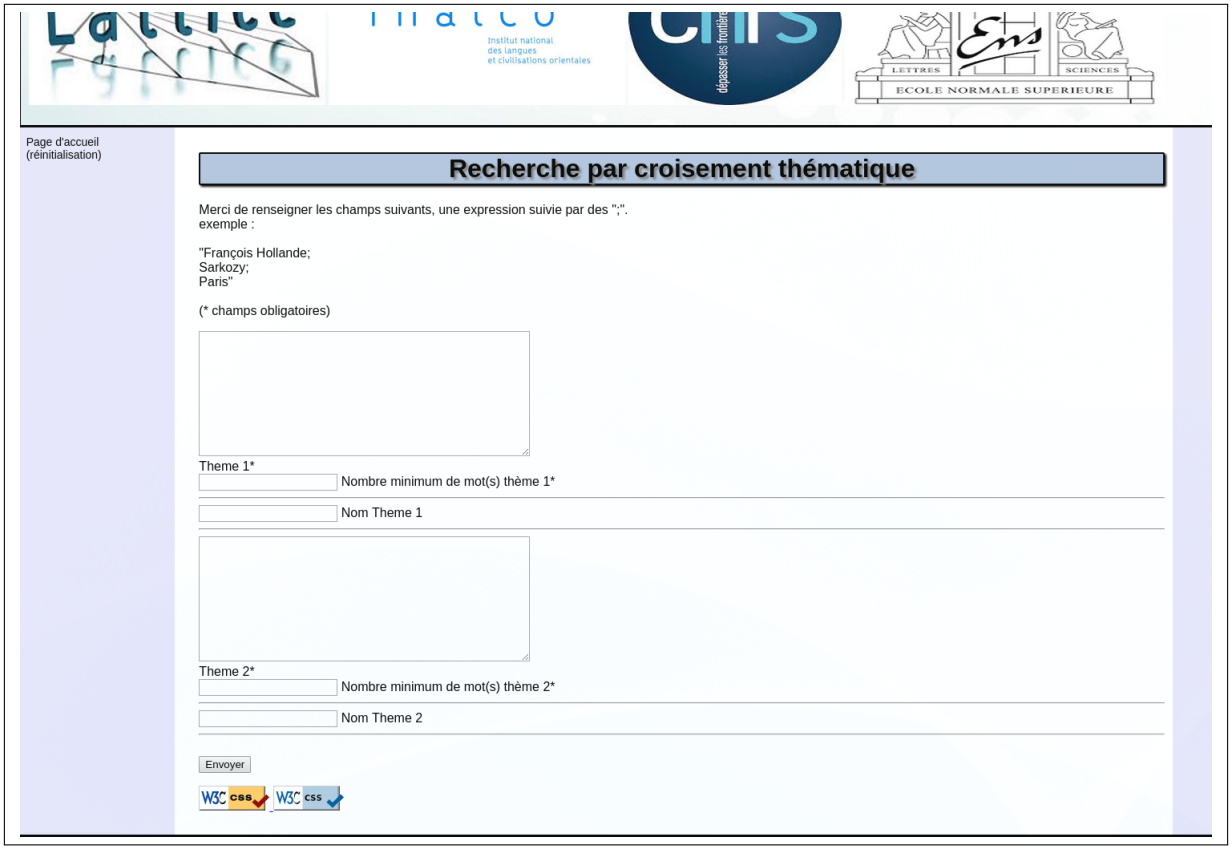


FIGURE 2.9 – Interface - Accueil

L'utilisateur peut ici :

— Définir une liste de mots-clés séparés par un point-virgule,
— Nommer cette liste de mots-clés correspondant à un thème du nom de ce thème,
— Définir le seuil du nombre de mots-clés que le système doit trouver dans un article (article+titre) pour juger que l'article en question contient le thème attendu par l'utilisateur et ajouter cet article à la liste de résultats

Les mêmes étapes sont attendues pour les deux thèmes, en les renseignant l'utilisateur peut avoir accès au résultat :

Sur GitHub https://github.com/PerlDancer/Dancer2
ou encore le site wikipédia https://en.wikipedia.org/wiki/Dancer_(software).
53. https://fr.wikipedia.org/wiki/Mod%C3%A8le-vue-contr%C3%B4leur

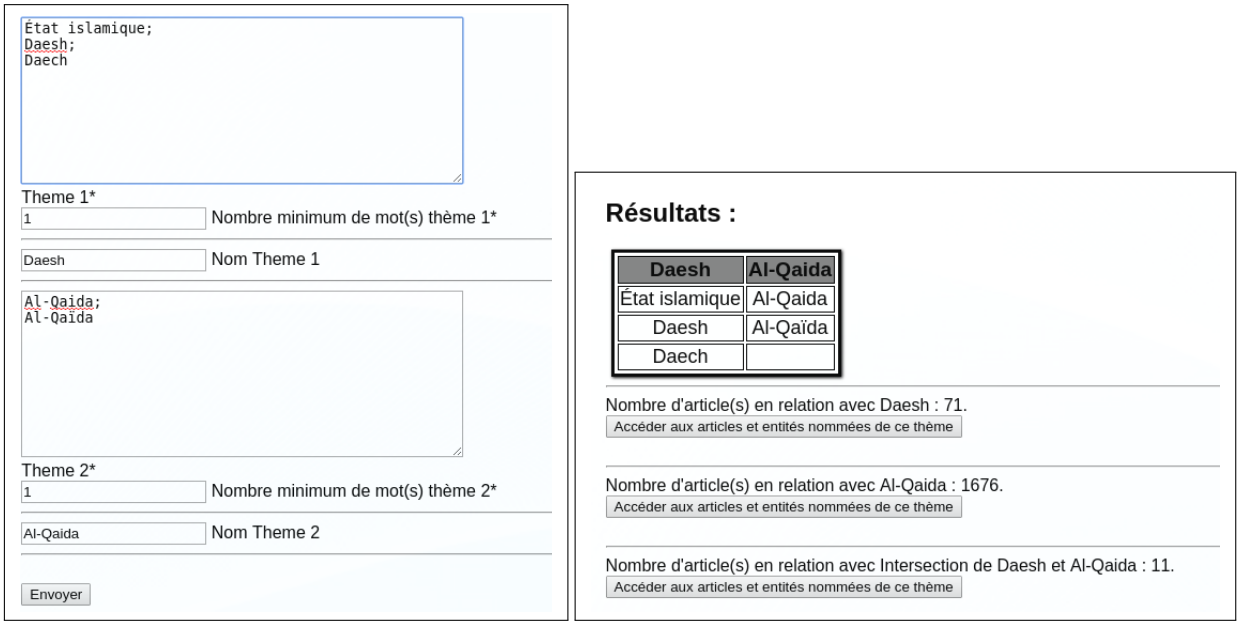


FIGURE 2.10 – Interface - Requêtes/Réponses

On observe ici à gauche la requête et à droite les résultats. L'utilisateur peut ensuite consulter chacun des thèmes individuellement, ou l'intersection des deux (pour plus de détail sur le code on consultera l'annexe C p.87). Cette étape peut être résumée dans le schéma suivant. Le programme Perl, prend en entrée la liste de mots-clefs et renvoie un index correspondant aux articles ayant satisfait les critères de l'utilisateur :

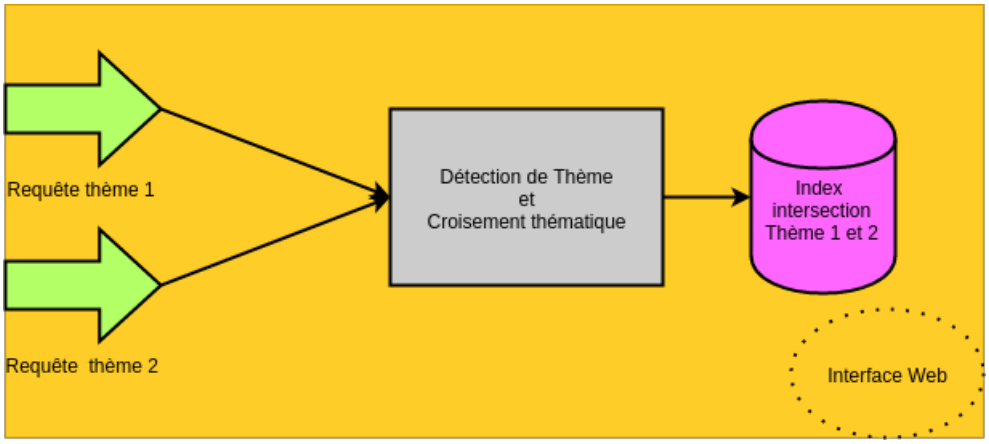


FIGURE 2.11 – Résumé de l'étape Requêtes/Réponses

L'utilisateur est ensuite libre d'explorer chacun des trois thèmes (thème 1, 2 ou l'intersection des deux). À l'affichage des résultats on peut observer les résultats suivants (en haut le thème 1, et en bas l'intersection des thèmes) :

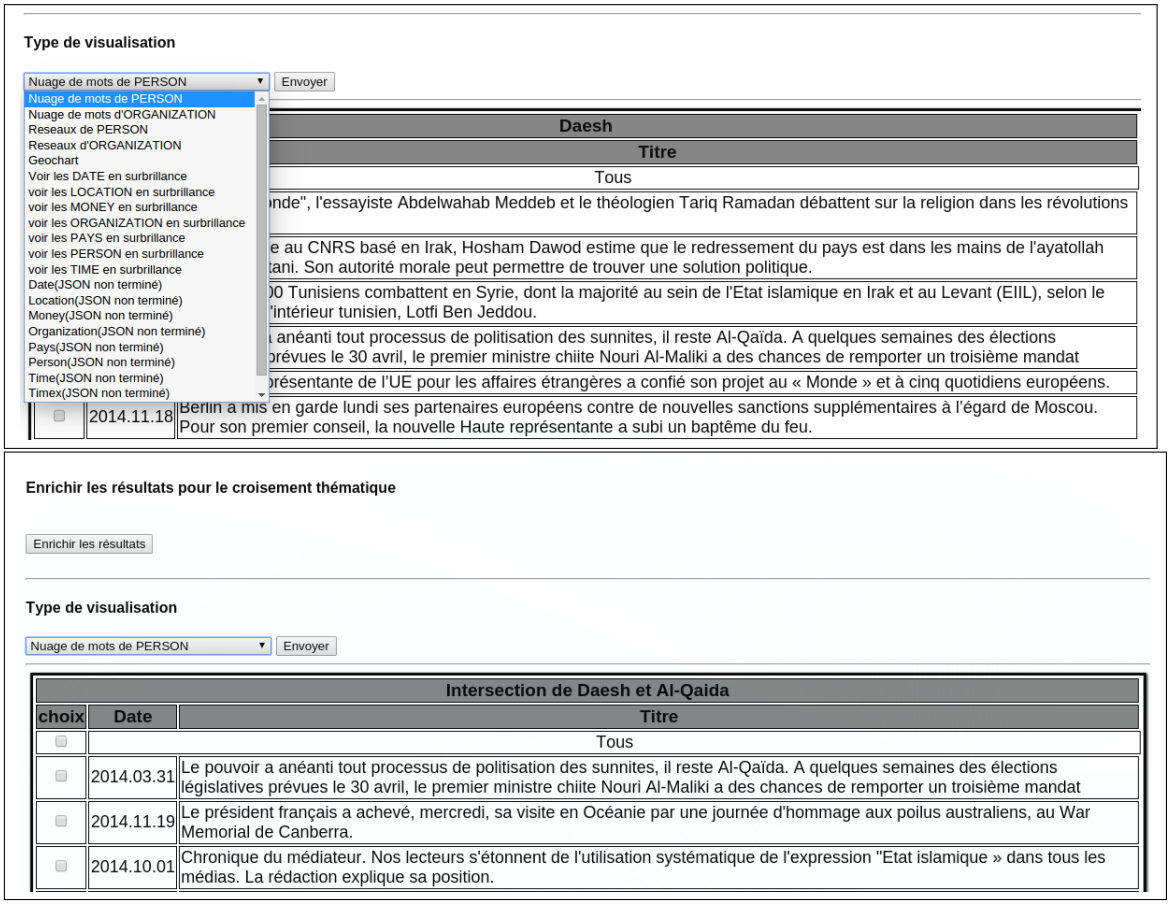


FIGURE 2.12 – Interface - Réponses Thème 1 / Intersection

Dans les deux cas on voit s'afficher la liste des titres d'articles correspondants au thème recherché, ainsi que leur date d'édition. Ces informations sont obtenues par l'intermédiaire d'un accès à la base de données en utilisant les modules DBI et $DBD :: SQLite$ [54] via des requêtes s'appuyant sur les *id* (voir l'annexe D p.89). L'utilisateur peut ensuite sélectionner les articles qui semblent correspondre à sa recherche et un type de visualisation. Nous verrons les types de visualisations proposés dans la section suivante. Auparavant parlons de l'aspect enrichissement par similarité que nous avions détaillé à la section 2.2.3 p.28.
En effet dans la figure 2.12 nous pouvions remarquer la présence d'une option supplémentaire pour l'intersection (image du bas), « Enrichir les résultats ».

En cliquant sur cette option, l'utilisateur accède à l'écran suivant :

54. Pour DBI : http://dbi.perl.org/,
Pour $DBD :: SQLite$: http://search.cpan.org/~msergeant/DBD-SQLite-0.31/lib/DBD/SQLite.pm.

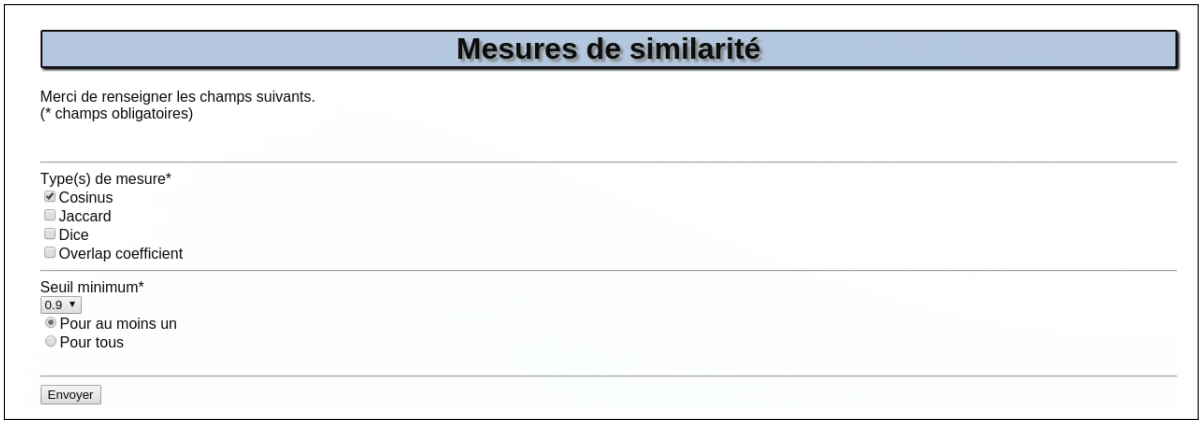


FIGURE 2.13 – Interface - Enrichissement par similarité

Il peut sur cet écran définir les paramètres de la mesure d'enrichissement qu'il souhaite :

A) le type de calcul

- Cosinus
- Dice
- Jaccard
- Overlap

B) le seuil d'exigence

- Un seuil paramétrable par dixième sur l'intervalle [0.1-0.9], renvoyant comme vrais les résultats supérieur à ce seuil

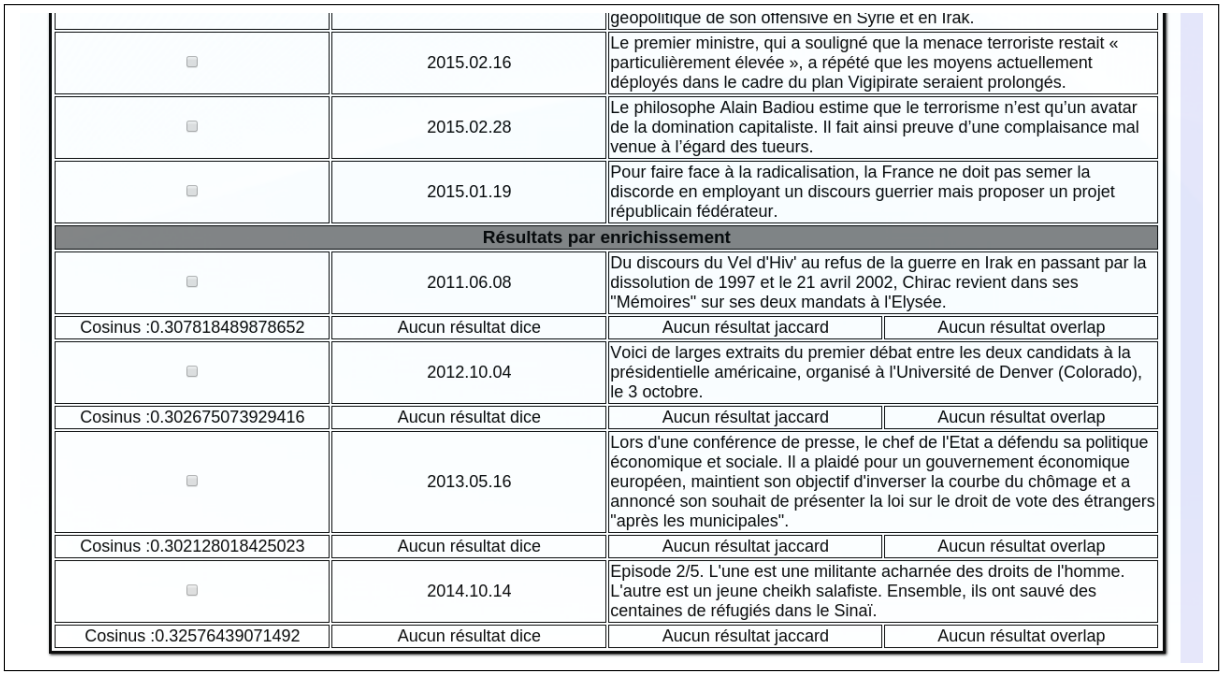
C) le degré d'exigence

- On considère qu'une seule valeur au dessus du seuil est suffisante ou l'ensemble des valeurs

Pour une requête sur le thème précédent avec comme paramètres :

- un type de calcul demandant une réponse avec Cosinus, Dice, Jaccard ;
- un seuil d'exigence supérieur à 0.3 ;
- et un degré d'exigence faible, soit un seul des trois types de calcul au dessus du seuil

L'utilisateur obtient en plus des 11 articles de son intersection thématique initiale, 4 résultats supplémentaires :



		géopolitique de son offensive en Syrie et en Irak.	
☐	2015.02.16	Le premier ministre, qui a souligné que la menace terroriste restait « particulièrement élevée », a répété que les moyens actuellement déployés dans le cadre du plan Vigipirate seraient prolongés.	
☐	2015.02.28	Le philosophe Alain Badiou estime que le terrorisme n'est qu'un avatar de la domination capitaliste. Il fait ainsi preuve d'une complaisance mal venue à l'égard des tueurs.	
☐	2015.01.19	Pour faire face à la radicalisation, la France ne doit pas semer la discorde en employant un discours guerrier mais proposer un projet républicain fédérateur.	
Résultats par enrichissement			
☐	2011.06.08	Du discours du Vel d'Hiv' au refus de la guerre en Irak en passant par la dissolution de 1997 et le 21 avril 2002, Chirac revient dans ses "Mémoires" sur ses deux mandats à l'Elysée.	
Cosinus :0.307818489878652	Aucun résultat dice	Aucun résultat jaccard	Aucun résultat overlap
☐	2012.10.04	Voici de larges extraits du premier débat entre les deux candidats à la présidentielle américaine, organisé à l'Université de Denver (Colorado), le 3 octobre.	
Cosinus :0.302675073929416	Aucun résultat dice	Aucun résultat jaccard	Aucun résultat overlap
☐	2013.05.16	Lors d'une conférence de presse, le chef de l'Etat a défendu sa politique économique et sociale. Il a plaidé pour un gouvernement économique européen, maintient son objectif d'inverser la courbe du chômage et a annoncé son souhait de présenter la loi sur le droit de vote des étrangers "après les municipales".	
Cosinus :0.302128018425023	Aucun résultat dice	Aucun résultat jaccard	Aucun résultat overlap
☐	2014.10.14	Episode 2/5. L'une est une militante acharnée des droits de l'homme. L'autre est un jeune cheikh salafiste. Ensemble, ils ont sauvé des centaines de réfugiés dans le Sinaï.	
Cosinus :0.32576439071492	Aucun résultat dice	Aucun résultat jaccard	Aucun résultat overlap

FIGURE 2.14 – Interface - Enrichissement par similarité : Résultats

Les résultats n'étant pas toujours pertinents, charge à l'utilisateur de bien définir les thèmes en amont et de faire plusieurs tests avec des paramètres différents. On trouvera un exemple intéressant dans le chapitre suivant (voir section 3.3 p.70).

L'implémentation du calcul peut être consulté dans l'annexe E p.91, l'étape, quant à elle, peut être résumée dans le schéma suivant :

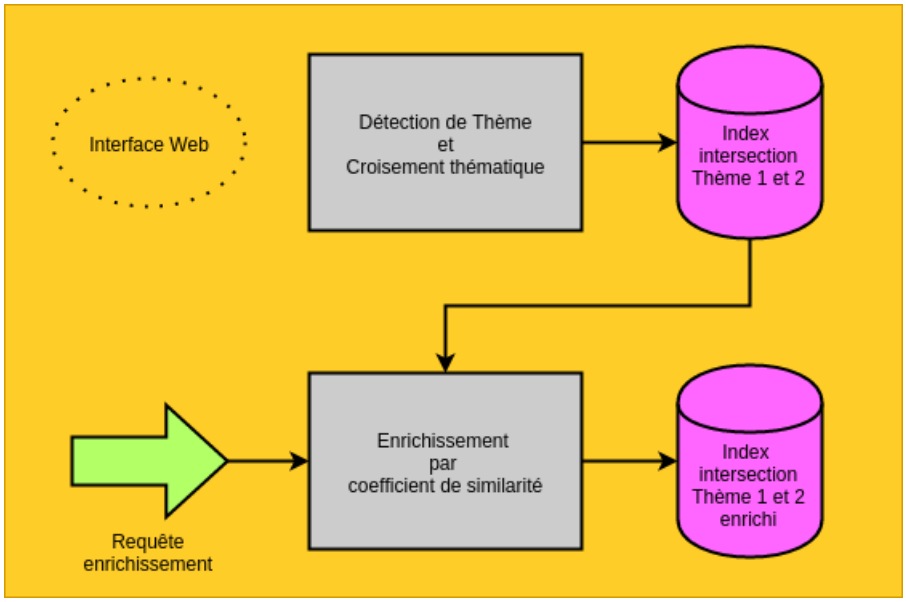


FIGURE 2.15 – Résumé de l'étape Requêtes/Réponses+Enrichissement

Nous abordons dans la section suivante les types de visualisations disponibles à partir d'une des quatre fenêtres de résultats (thème 1, thème 2, intersection, ou intersection enrichi).

2.4.3 Visualisation

L'étape de visualisation, selon que l'on part d'un thème simple (thème 1, thème 2 ou intersection des deux) ou d'un thème enrichi (par coefficient de similarité sur l'intersection des thèmes 1 et 2) peut être envisagée selon le schéma suivant :

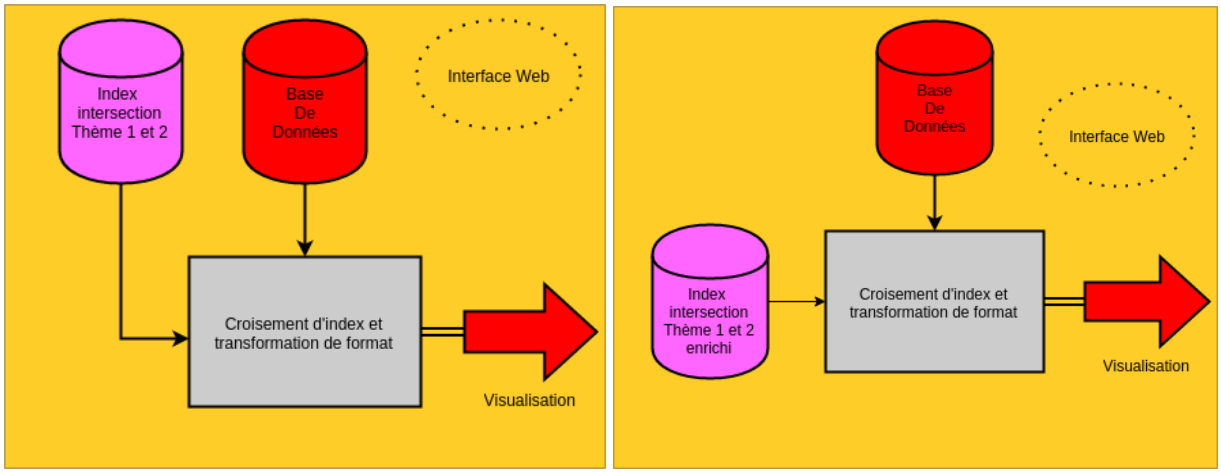


FIGURE 2.16 – Résumé de l'étape Intersection des index et de la base de données, puis visualisation

L'index fourni en entrée nous permet de faire le lien entre thème et articles, et par extension avec les EN contenus dans ces articles reste à représenter ces EN.

En observant certains outils de visualisation en TAL comme cortext [55] et en le testant sur notre corpus, on obtient une première idée de l'interface que l'on souhaiterait obtenir. On pourra trouver en annexe G p.99 certains résultats des tests effectués sur notre corpus. Cependant à la différence de cet outil nous souhaiterions une visualisation avec consultation des données de manière dynamique, et livrant plus d'informations sémantiques.

La consultation d'écrits sur la visualisation, avec notamment l'excellent ouvrage de Nathan Yau [Yau, 2013], spécialiste de la représentation de données et auteur du site *flowingdata* [56] nous offre un panorama des techniques de visualisations à envisager.

Assez tôt, on se tourne vers le javascript comme langage support pour nos visualisations. Par la facilité avec laquelle il permet de produire une visualisation à partir de données, et aussi par sa flexibilité (assuré par des bibliothèques comme JQuery [57], il semble le langage à privilégier par excellence.

Ainsi lors de l'interrogation de notre base de données on obtient en retour un identifiant et une liste de pays :

```
sqlite> SELECT id, en FROM PAYS WHERE(id = '2014.3387');
2014.3387|Belgique;Etats-Unis;France;Jordanie;Liban;Qatar;Turquie
```

Ce type de requête est systématisé dans nos scripts Perl comme nous avons pu le dire auparavant [58]. En retour on obtient une chaîne de caractère associée à un identifiant.

55. http://www.cortext.net/
56. http://flowingdata.com/
57. https://fr.wikipedia.org/wiki/JQuery
58. On consultera à nouveau l'annexe D p.89 pour les détails de l'interrogation des base de données par scripts.

La chaîne de caractère est ensuite découpée et, avec son id, transformée en JSON (ou JavaScript Object Notation [59]) modèle de données pour javascript comme par exemple dans le résultat de requête suivant :


```
{
   "2014.3387" : [
      "Belgique;Etats-Unis;France;Jordanie;Liban;Qatar;Turquie"
   ],
   "2014.18766" : [
      "Chine;France"
   ],
   "2012.23977" : [
      "Afghanistan;Canada;Chine;Etats-Unis"
   ],
   "2015.1428" : [
      ""
   ],
   "2014.15684" : [
      "Afghanistan;États-Unis"
   ],
   "2014.19512" : [
      "Etats-Unis;France"
   ],
   "2011.13356" : [
      "Afghanistan;France;Japon;Réunion;États-Unis"
   ],
   "2013.18471" : [
      "Afghanistan;Chine;France;Japon;Jordanie;Liban;Mali"
   ],
   "2014.22018" : [
      "Algérie;Colombie;Espagne;Finlande;France;Japon;Jordanie;Maroc;Suisse;Tu
   ],
   "2015.2667" : [
      "France"
   ],
   "2014.19510" : [
      "Israël"
   ],
   "2014.23066" : [
      "Etats-Unis;France"
   ],
   "2014.20518" : [
      "Allemagne;Belgique;Djibouti;France;Israël;Italie;Soudan;Suède;Yémen"
   ],
   "2015.1158" : [
      "France;Israël;Qatar;Turquie"
   ],
   "2014.22893" : [
```


59. https://fr.wikipedia.org/wiki/JavaScript_Object_Notation

```
        "Algérie;Mali"
    ]
}
```

Il nous reste à transformer ces objets javascript selon nos besoins, en croisant les données, en essayant de garder à l'esprit les besoins utilisateurs auxquelles nous souhaitions répondre dans l'introduction (voir 0.1 p.11).

Timeline

À minima on souhaite produire un retour aux articles du thème, classés par date. On souhaite aussi que l'utilisateur puisse avoir la possibilité de naviguer parmi les articles de façon à pouvoir trouver l'information qu'il jugera pertinente. Pour cela nous adaptons le code de la librairie timeliner.js [60] A partir du retour au texte, et avec une modification minime du code on peut proposer à l'utilisateur de visualiser les autres EN en surbrillance dans le texte de façon à faciliter l'accès à l'information recherchée.

FIGURE 2.17 – Timeline des articles d'un thème, mode plein texte et visualisation d'EN *person*

De cette façon l'utilisateur a accès à tous les types d'EN que nous avons pu extraire, selon un modèle temporel et en contexte.

60. Et plus particulièrement celui-ci https://github.com/technotarek/timeliner .

Geochart

Outre l'accès au texte selon l'aspect temporel on souhaite fournir à l'utilisateur des informations selon un cadre spatial. Pour cela on tire partie de l'extraction par règles des EN *pays* croisée avec la librairie de Google Geochart [61]. Cette librairie fonctionne avec les noms de pays en langue anglaise. Nous convertissons les noms et les transformons en JSON avant de pouvoir les afficher selon le modèle suivant :

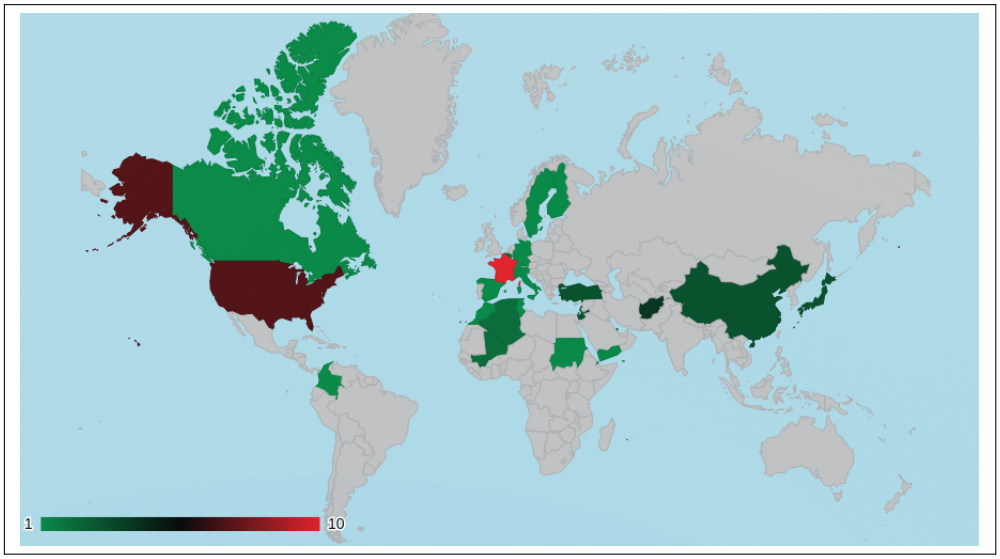


FIGURE 2.18 – Pays dont les articles du thème font mention

Dans cet exemple on constate que le pays le plus représenté dans ce thème est cité dans 10 articles. En survolant la carte des pays l'utilisateur peut voir apparaître les noms de pays et le nombre d'articles dans lesquels ils sont cités. Les couleurs dépendent du nombre de fois où le pays est cité dans le thème. Si le pays n'est pas cité dans le thème il apparaît en gris. Le(s) pays le(s) plus cité(s) apparaissent en rouge, le(s) moins cité(s) apparaissent en vert. Entre les deux une frise couleur nous permet de déterminer où les autres pays se situent en nombre de citation. Au survol d'un pays cité dans le thème, un curseur s'affiche sur la frise pour situer celui-ci par rapport aux autres.

D3

Enfin une librairie très riche en visualisations et représentation des données est la librairie *Data Driven Document* ou *D3* [62]. Elle est basée sur la manipulation du *Document Objet Modèle* ou *DOM* [63] avec *HTML*, *CSS* et *SVG*. Elle est aussi très facile d'accès et libre de droit, elle permet en outre de manipuler les données selon différents angles ce qui correspond à la perspective de personnalisation des données que nous recherchions.

Nous utilisons ensuite les données transformées au format *JSON* dans différentes implémentations de D3. Une fois réflexion faite sur le type de visualisation qui pourrait être informative pour l'utilisateur, nos critères de sélection des librairies

61. https://developers.google.com/chart/interactive/docs/gallery/geochart
62. http://d3js.org/
63. https://fr.wikipedia.org/wiki/Document_Object_Model

se bornent à la facilité d'implémentation et la popularité de la librairie sur D3.

Nuage de mots

On considère que la fréquence d'apparition d'une EN est représentative de son importance. Selon cette idée et toujours dans une optique de représentation des données on estime qu'un accès aux données via un nuage de mots pourrait être informatif pour l'utilisateur. Ainsi en un coup d'œil l'utilisateur peut déterminer quelle est l'entité ou quelles sont les entités prégnante(s) du thème. On utilise à nouveau une sélection de D3 [64] pour créer ces nuages de mots.

On peut ensuite observer ce type de résultat :

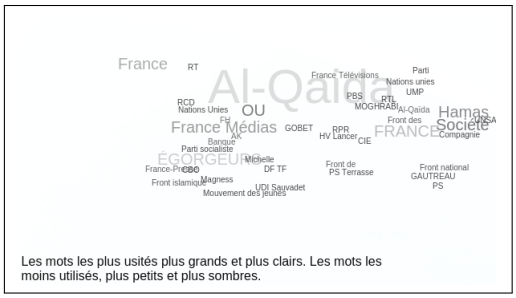


FIGURE 2.19 – Visualisation de type nuage de mots d'organisations

On constate par exemple, dans le nuage de mots ci-dessus, la surreprésentation dans ce thème d'« Al-Qaida » et de « Hamas »

FIGURE 2.20 – Visualisation de type nuage de mots de personnes

« François Hollande » est ici le plus représenté.

En complément de ce nuage de mot qui rend l'information accessible rapidement à l'utilisateur, nous aurions souhaité produire un histogramme de fréquence, par manque de temps nous avons abandonné cet aspect.

Reseaux

Il nous apparaît que visualiser les interconnexions entre acteurs d'une même thématique peut être un élément pertinent dans l'apport d'information pour l'utilisateur. Nous cherchons donc à établir un visualisation en réseaux de *person* et *organization*. Pour cela on utilise plusieurs bibliothèques [65] D3 que nous implémentons selon nos besoin. On peut ensuite observer ce type de résultat :

64. http://bl.ocks.org/ericcoopey/6382449

65. On pourra consulter notamment : http://christophergandrud.github.io/d3Network/ ou encore http://bl.ocks.org/mbostock/1153292.

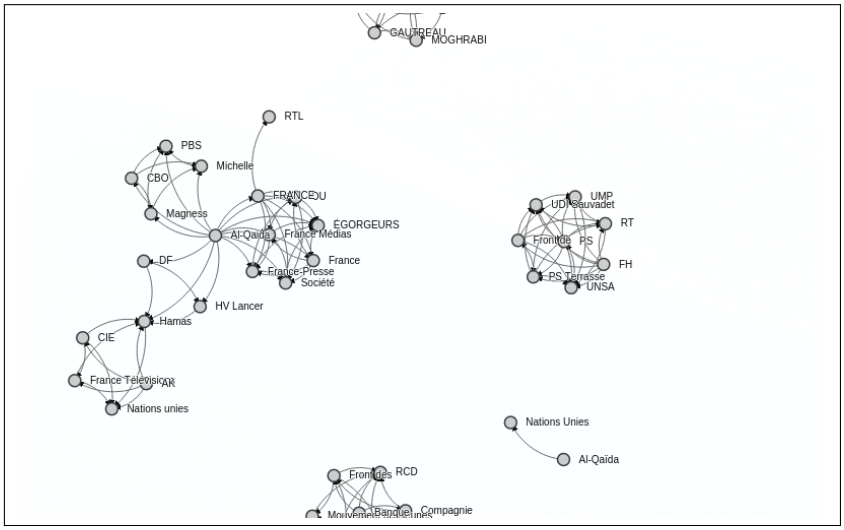


FIGURE 2.21 – Visualisation en réseaux d'organisations

Dans ce cas on peut remarquer des « galaxies » déconnectés les unes des autres (le thème n'est pas unifié entre article sur ces EN), avec certains faisceaux de convergence (comme ici avec « Al-Qaida » et « Hamas »)

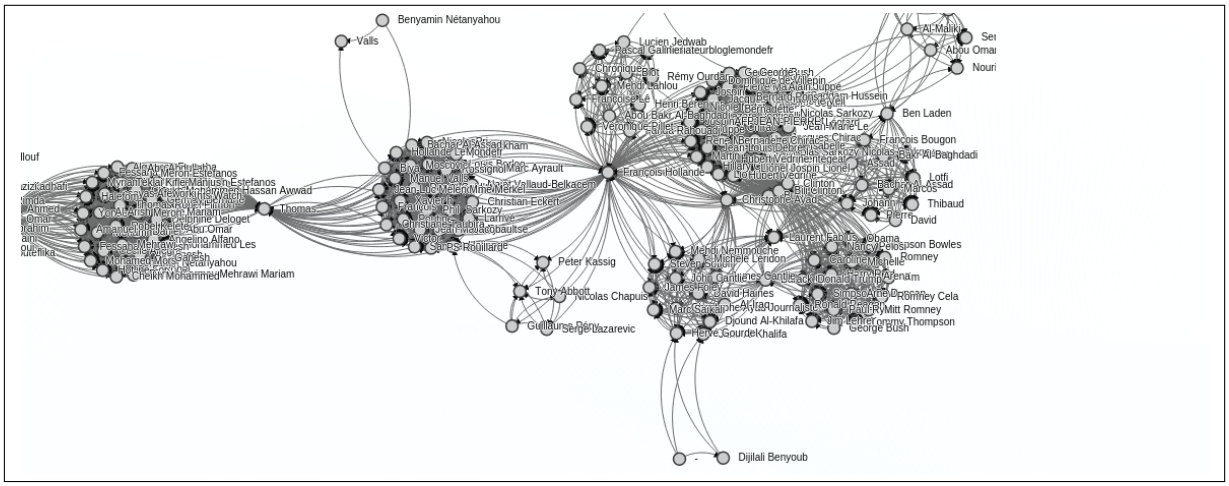


FIGURE 2.22 – Visualisation en réseaux de personnes

A l'inverse on peut observer dans le cas ci-dessus, une unification globale du thème avec certains faisceaux de convergence sur des éléments clés (comme ici avec « François Hollande », « Laurent Fabius » ou encore « Ben Laden »).

Ces deux exemples sont assez caractéristiques des visualisations en réseaux de type personnes ou organisation, les premières sont souvent denses dans l'affichage des données, les secondes montrent souvent une « déconnexion » entre EN. Il convient donc, selon les cas, de sélectionner minutieusement les articles pour ne pas être confronté à un rendu visuel saturé en informations ou d'orienter la recherche dans un cadre précis de façon à établir des connexions pertinentes entre acteurs.

D'un point de vue technique on remarque que les librairies permettent de déplacer dynamiquement les galaxies d'EN, ce procédé manque néanmoins de facilité face à un grand nombre de données et il conviendrait de l'améliorer.

En observant conjointement Réseaux et Nuages de mots on peut définir quels entités sont caractéristiques : ici, on constate que « François Hollande », « Hamas » et « Al-Qaida » sont surreprésentés dans les articles de presse pour ce thème. Au regard de la période (via Timeline section 2.17 p.58) , on peut définir que cette surreprésentation s'effectue sur une période allant du 31/03/14 au 28/02/15 (on pourrait aussi dater chaque article). Enfin les pays concernés sont visualisables sur GeoChart (section 2.18 p.59). Tout ces éléments sont des points caractéristiques du thème abordé dans la partie 2.4.2 p.50 à savoir l'intersection des thèmes « Al-Qaida » et « Daesh ».

2.5 Conclusion

Nous avons vu lors de ce chapitre et du précédent les éléments nécessaires à l'établissement de notre chaîne de traitement. On visualisera l'ensemble de celle-ci dans le schéma suivant :

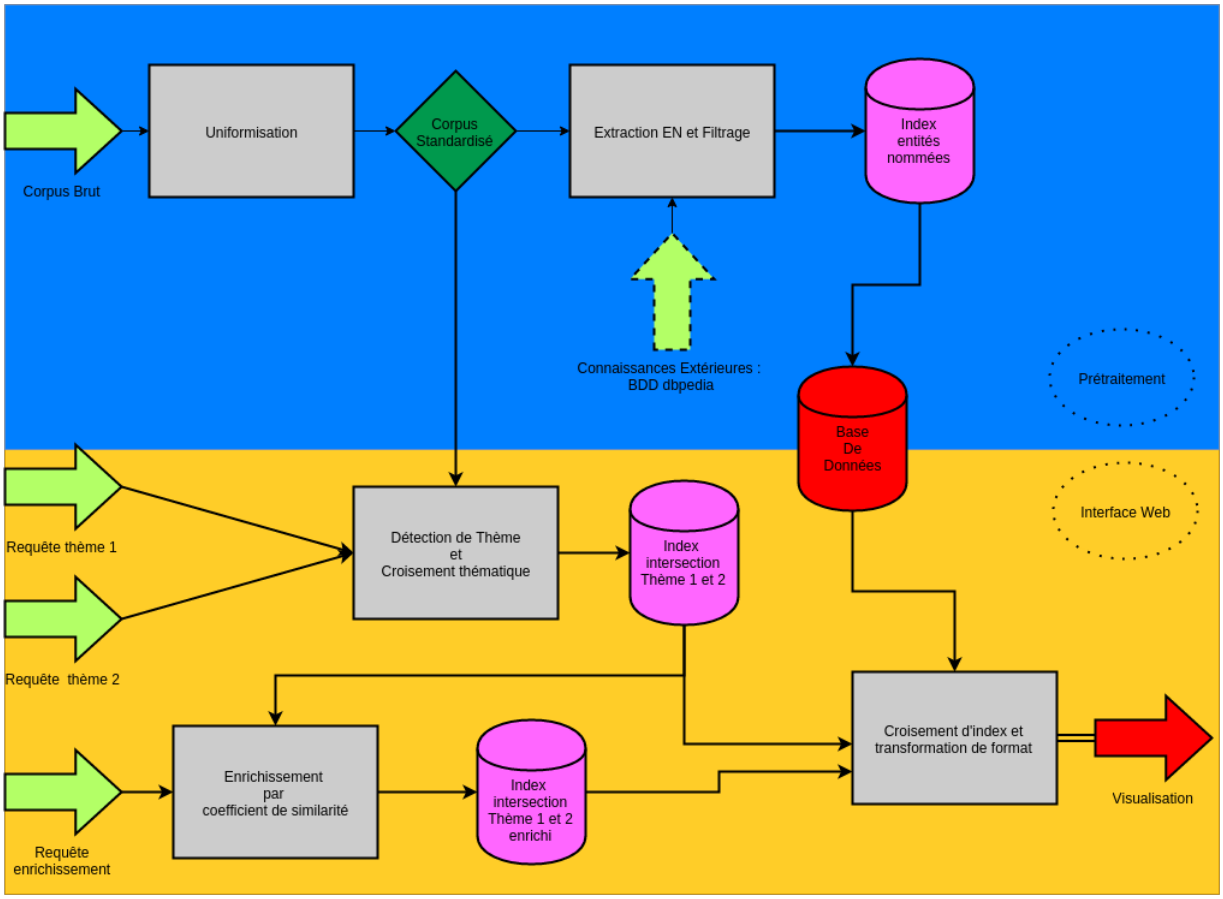


FIGURE 2.23 – Filtrage sémantique et visualisation de données textuelles

CHAPITRE 3

RÉSULTATS

Sommaire

3.1 Introduction

Dans ce chapitre nous aborderons essentiellement la question de la visualisation. Nous verrons l'exemple de réponses à deux requêtes utilisateurs. Le première requête concernera un thème avec un grand nombre d'articles en réponse. La seconde sur l'exploitation d'informations en rapport avec un thème ayant peu d'article en réponse à la requête utilisateur.

3.2 Requête 1 : nombre élevé d'informations en réponse

Comme première démonstration de l'utilisation de l'outil, nous essayons d'abord de récolter de l'information en considérant l'idée d'explorer une thématique pouvant occasionner un grand nombre de résultats, puis de réduire peu à peu cette masse d'information jusqu'à faire émerger un axe qui pourrait répondre à une problématique. On considère ainsi la question suivante : En quoi l'utilisation de drones influence-t-elle le terrorisme ?

On imagine à priori que l'intersection des deux thèmes que représentent les « drones » et le « terrorisme » nous donnera au moins une centaine d'articles correspondants. Un premier test avec comme requêtes les notions entrées tel quel semble nous donner raison :

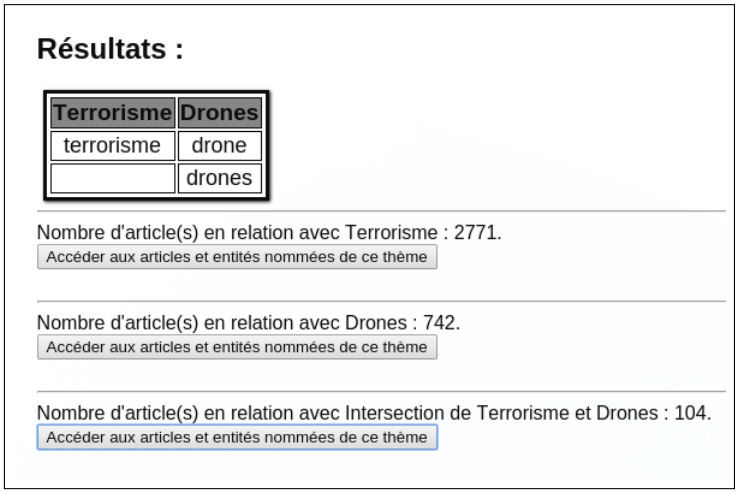


FIGURE 3.1 – Nombre d'articles en réponse à une requête simple « *Terrorisme* » et « *Drone* »

Nous définissons les caractéristiques de notre thème « *terrorisme* » en utilisant la liste de mots-clés en relation avec le terrorisme issue de des outils de textométrie mais simplifié de 82 items à 30. On obtient alors les résultats suivants :

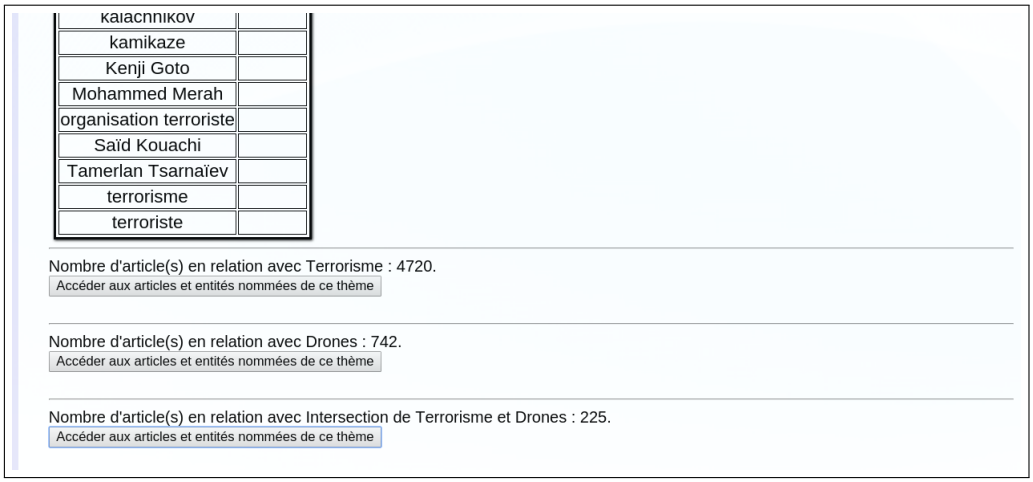


FIGURE 3.2 – Nombre d'articles en réponse à une requête avec vocabulaire spécifique « *Terrorisme* » et « *Drone* »

On augmente le nombre d'articles trouvés en rapport avec notre thème. Cependant si le thème est plus précis, le volume des données rend l'exploitation de certaines visualisations difficile, par exemple :

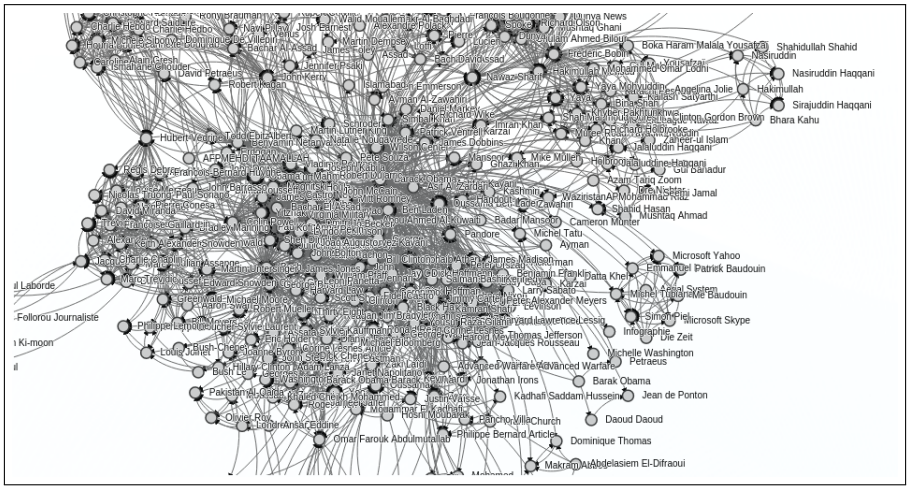


FIGURE 3.3 – Saturation d'informations pour une visualisation en réseaux d'*EN person* pour l'intersection des thèmes « *Terrorisme* » et « *Drone* »

Où l'on constate que la masse d'EN de type *person* ne permet pas d'exploiter des informations pertinentes. Les organisations étant moins présentes dans les articles, leur visualisation semble plus pertinente :

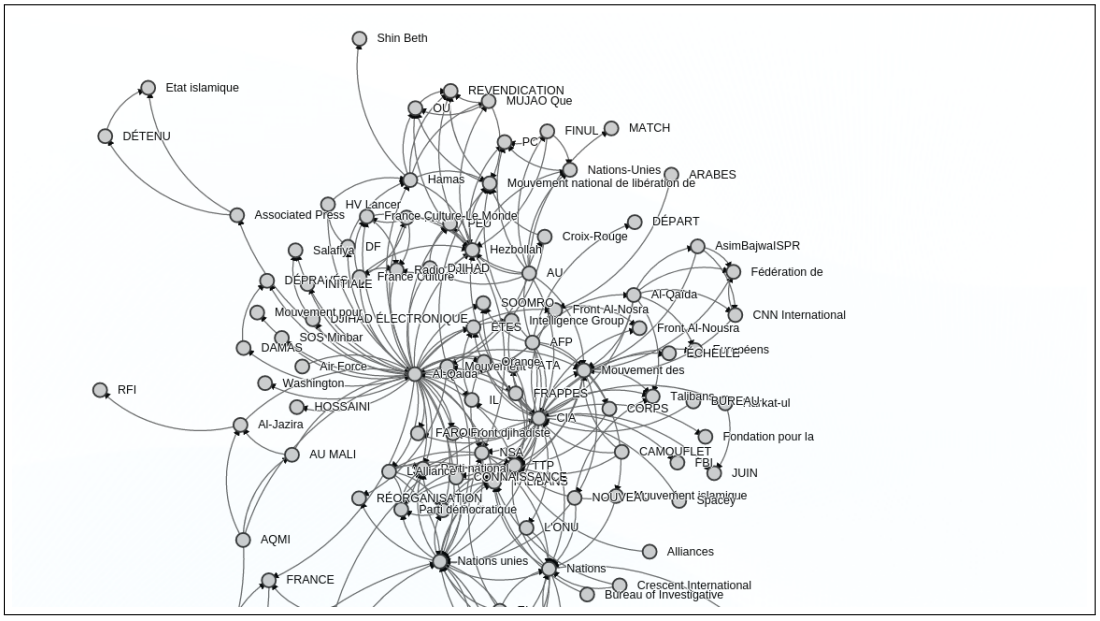


FIGURE 3.4 – Visualisation en réseaux d'*EN organization* pour l'intersection des thèmes « *Terrorisme* » et « *Drone* »

On constate en effet la présence de certains pôles d'attraction comme *Al-Qaida*, *CIA*, ou encore *Nations Unies*. On peut aussi observer la fréquence des mêmes EN en nuage de mots :

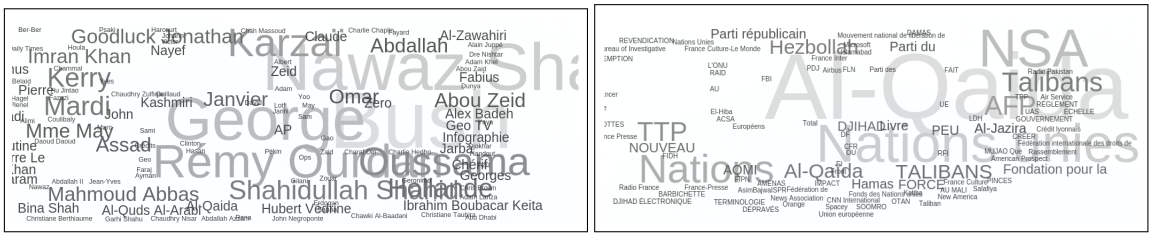


FIGURE 3.5 – Visualisation en réseaux d'*EN organization* pour l'intersection des thèmes « *Terrorisme* » et « *Drone* »

Enfin le traitement par pays nous permet une visualisation qui nous donne une idée des pays concernés par le croisement de ces deux thématiques.

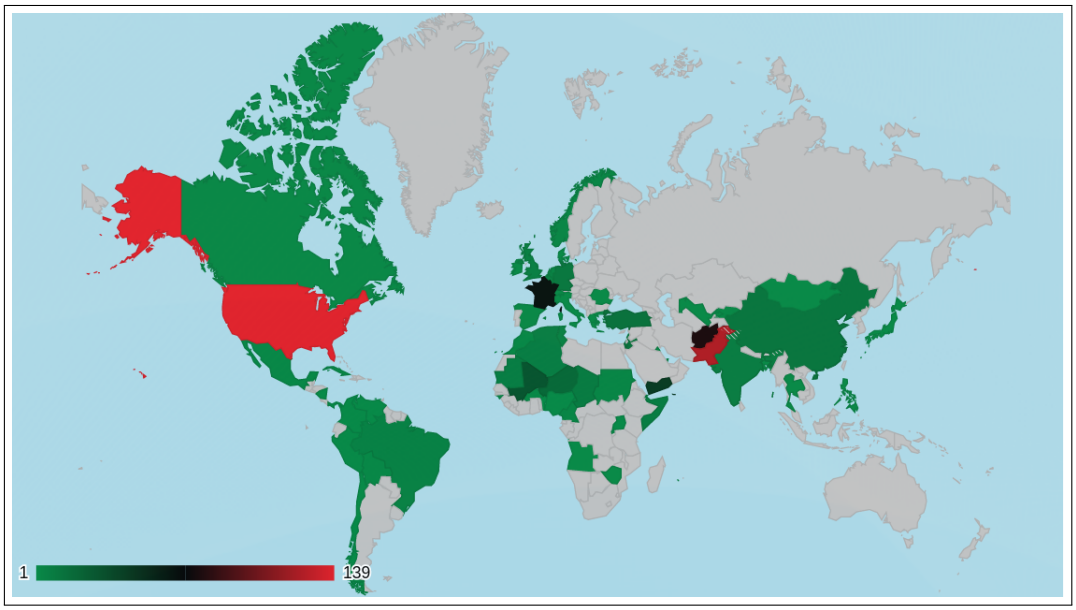


FIGURE 3.6 – Pays concernés par l'intersection des thèmes « *Terrorisme* » et « *Drone* » entre 2009.01 et 2015.02

La forte représentation d'« Al-Qaida » par fréquence (nuage de mots) est comme point central de plusieurs réseaux d'EN, on décide de simplifier la thématique terrorisme en regardant uniquement l'interaction entre « drones » et « Al-Qaida » [1].

1. Il conviendrait, dans l'optique d'une étude complète, de faire cette opération à plusieurs reprises et selon des axes différents.

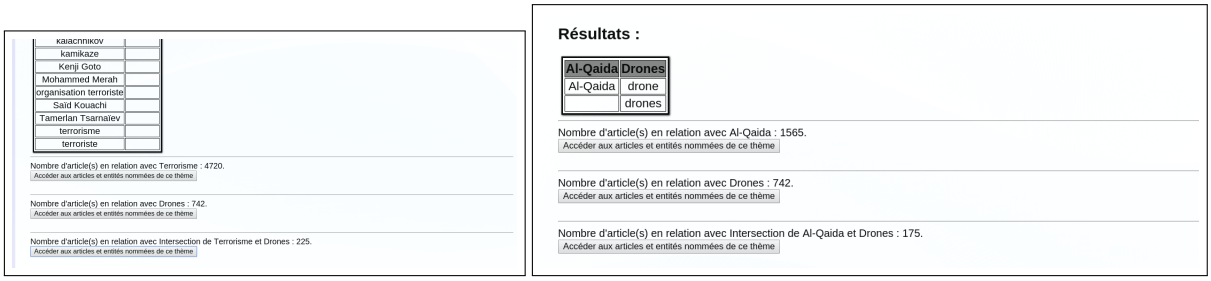
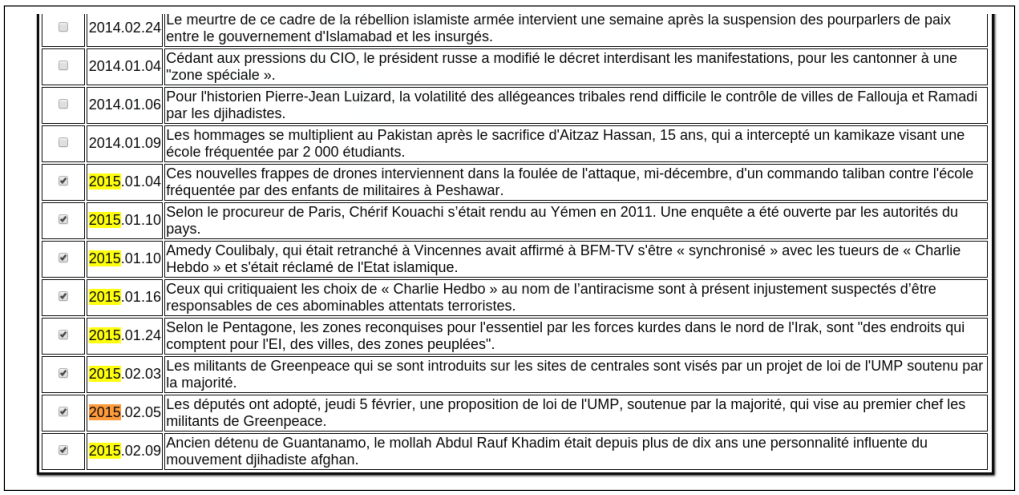


FIGURE 3.7 – Diminution du nombre de résultat en n'utilisant qu'un des axes contenu dans les résultats initiaux

On obtient un nombre inférieur de résultats pour les intersections (à droite) mais, malgré tout, encore trop pour les observer dans le détail. On décide d'agir sur la temporalité et de ne considérer que les articles de 2015 soit uniquement les huit derniers résultats :



☐	2014.02.24	Le meurtre de ce cadre de la rébellion islamiste armée intervient une semaine après la suspension des pourparlers de paix entre le gouvernement d'Islamabad et les insurgés.
☐	2014.01.04	Cédant aux pressions du CIO, le président russe a modifié le décret interdisant les manifestations, pour les cantonner à une "zone spéciale ».
☐	2014.01.06	Pour l'historien Pierre-Jean Luizard, la volatilité des allégeances tribales rend difficile le contrôle de villes de Fallouja et Ramadi par les djihadistes.
☐	2014.01.09	Les hommages se multiplient au Pakistan après le sacrifice d'Aitzaz Hassan, 15 ans, qui a intercepté un kamikaze visant une école fréquentée par 2 000 étudiants.
☑	2015.01.04	Ces nouvelles frappes de drones interviennent dans la foulée de l'attaque, mi-décembre, d'un commando taliban contre l'école fréquentée par des enfants de militaires à Peshawar.
☑	2015.01.10	Selon le procureur de Paris, Chérif Kouachi s'était rendu au Yémen en 2011. Une enquête a été ouverte par les autorités du pays.
☑	2015.01.10	Amedy Coulibaly, qui était retranché à Vincennes avait affirmé à BFM-TV s'être « synchronisé » avec les tueurs de « Charlie Hebdo » et s'était réclamé de l'Etat islamique.
☑	2015.01.16	Ceux qui critiquaient les choix de « Charlie Hedbo » au nom de l'antiracisme sont à présent injustement suspectés d'être responsables de ces abominables attentats terroristes.
☑	2015.01.24	Selon le Pentagone, les zones reconquises pour l'essentiel par les forces kurdes dans le nord de l'Irak, sont "des endroits qui comptent pour l'EI, des villes, des zones peuplées".
☑	2015.02.03	Les militants de Greenpeace qui se sont introduits sur les sites de centrales sont visés par un projet de loi de l'UMP soutenu par la majorité.
☑	2015.02.05	Les députés ont adopté, jeudi 5 février, une proposition de loi de l'UMP, soutenue par la majorité, qui vise au premier chef les militants de Greenpeace.
☑	2015.02.09	Ancien détenu de Guantanamo, le mollah Abdul Rauf Khadim était depuis plus de dix ans une personnalité influente du mouvement djihadiste afghan.

FIGURE 3.8 – Sélection de huit résultats en 2015 pour l'intersection des thèmes « *Al-Qaida* » et « *Drone* »

On peut observer les évolutions géographiques entre janvier 2009 et février 2015 ou uniquement pour les deux premiers mois de 2015.

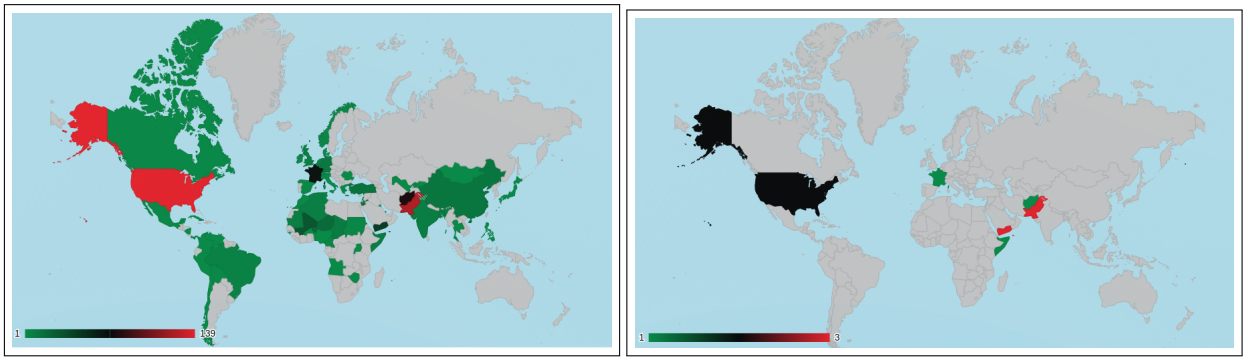


FIGURE 3.9 – Comparatif pour pays concernés par l'intersection des thèmes « *Al-Qaida* » et « *Drone* » entre 2009.01 et 2015.02 (gauche) et uniquement deux mois de 2015 (droite)

On observe des constantes comme les États-Unis, la France, l'Afghanistan, le Pakistan, le Yemen, et la Somalie. Cela nous apporte néanmoins peu d'information. La visualisation en réseaux montre curieusement un rapport entre notre thématique et les attentats de *Charlie Hebdo* début 2015 :

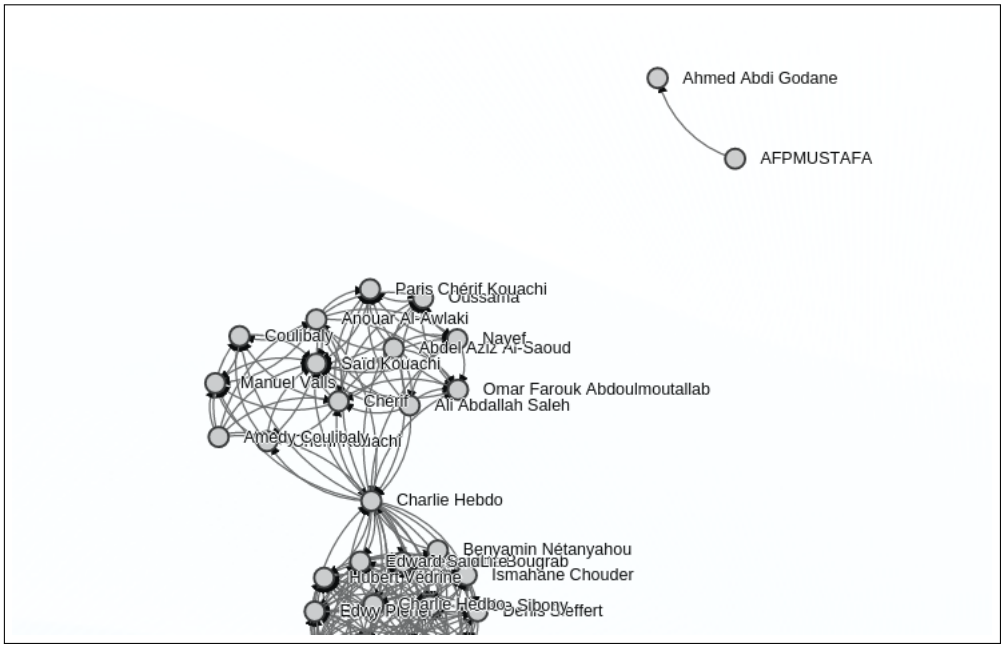


FIGURE 3.10 – Relation entre les attentats de « *Charlie Hebdo* » et l'intersection thématique « *Al-Qaida* » et « *Drone* »

Cette relation s'explique grâce à un retour au texte :

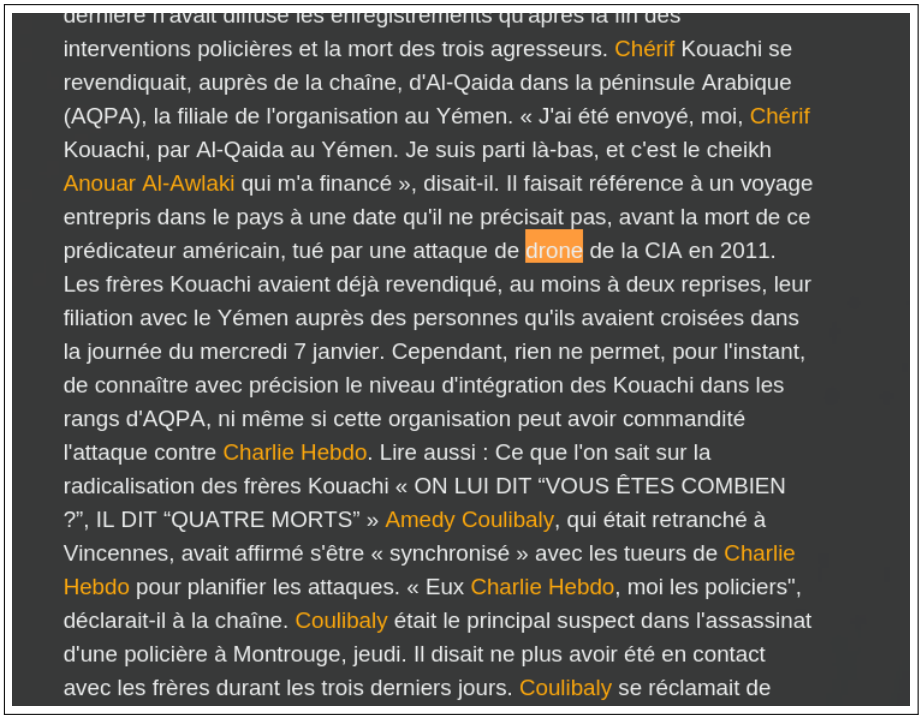


FIGURE 3.11 – Explication de la relation entre les attentats de « *Charlie Hebdo* » et l'intersection thématique

Où l'on constate que le commanditaire des terroristes ayant perpétré l'attentat de Charlie Hebdo a été tué par une attaque de drone. Une autre observation de texte fait émerger un point de vue intéressant :

FIGURE 3.12 – Influence de l'usage de drones sur l'émergence de nouveaux terroristes

Cette théorie partagé par le journal *The Guardian* dès 2012[2] selon laquelle « *L'usage indiscriminé de drones au Moyen-Orient cause beaucoup trop de morts civils, avertit un ex-dirigeant du contreterrorisme de la CIA.* ».

Nous estimons avoir suffisamment poussé l'expérience et nous arrêtons sur ces informations récoltées.

3.3 Requête 2 : nombre faible d'informations en réponse

Alors que dans la section précédente nous cherchions à diminuer et caractériser l'information par étape successive, nous chercherons ici à montrer l'utilité d'un enrichissement par coefficient de similarité sur un thème pour lequel la requête occasionne peu de résultat.

Dans cet exemple l'utilisateur souhaite savoir en quoi la question du « patrimoine de l'humanité » est affectée par le « terrorisme ».

Une première requête simplifiée avec les titres des deux notions que l'on va chercher à croiser montre la difficulté de trouver des articles partageant les deux notions et la nécessité de définir précisément les thèmes :

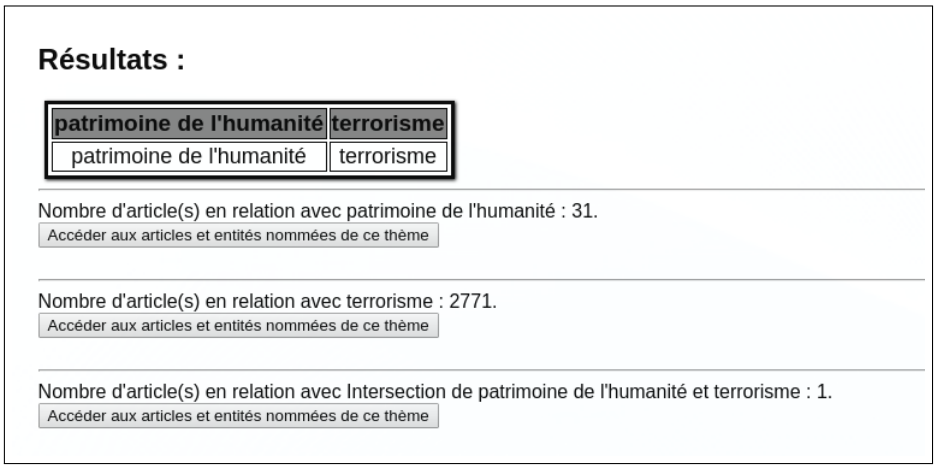


FIGURE 3.13 – Nombre d'article en réponse à la requête « *Terrorisme* » et « *Patrimoine de l'humanité* »

On constate en effet seulement 31 articles contenant le terme « *patrimoine de l'humanité* contre 2771 contenant le terme « *terrorisme* » sur 74 mois. Si l'on pouvait s'attendre à un tel écart, la proportion n'en est pas moins saisissante. De plus l'exploration textuelle sur le résultat partagé se montre tout aussi décevante du point de vue du croisement thématique :

2. voir l'article : http://www.theguardian.com/world/2012/jun/05/al-qaida-drone-attacks-too-broad, dernière consultation le 01/11/2015.

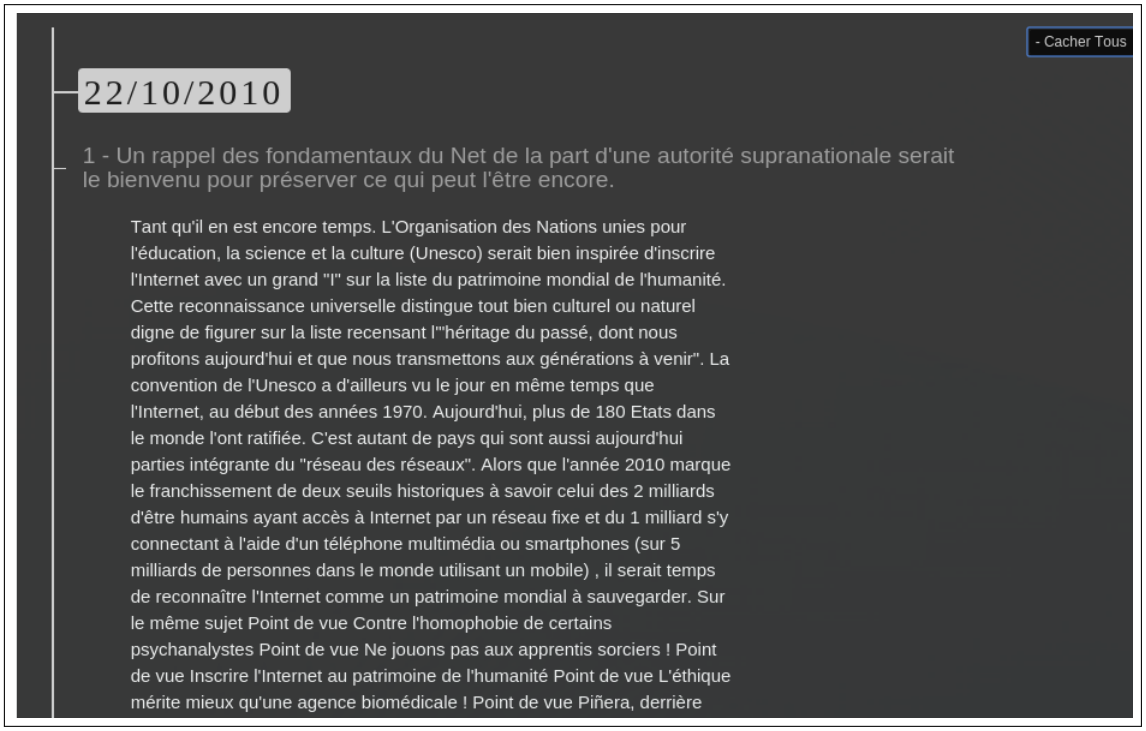


FIGURE 3.14 – Exploration textuelle du résultat commun à la requête « *Terrorisme* » et « *Patrimoine de l'humanité* »

Le résultat de ce texte n'a rien à voir avec la thématique réellement recherchée, il y est plus question de la sauvegarde de l'Internet. L'apparition dans ce texte du terme "terrorisme" est plus lié à un effet d'emphase de l'auteur cherchant à préserver les libertés individuelles que la description d'un acte lié au terrorisme.

Nous décidons d'éplucher les titres des articles ayant un résultat lié au mot terrorisme et de faire de même avec les articles traitant de patrimoine de l'humanité. Nous remarquons qu'à cette étape, l'utilisateur expert n'aurait aucune difficulté à rechercher certains mots-clés dans les titres d'articles pour sélectionner ceux qui conviendrait. Assez rapidement on arrive à sélectionner des articles pertinents :

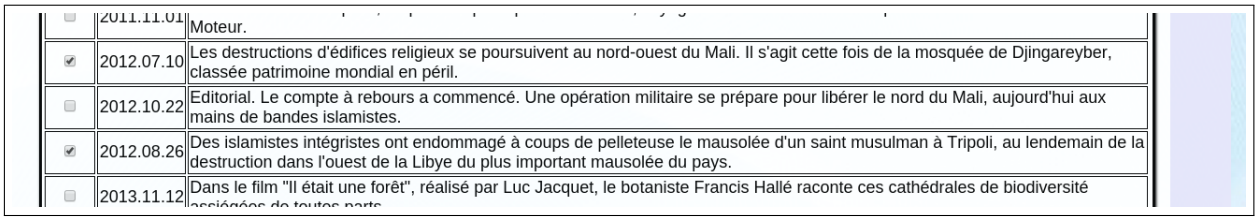


FIGURE 3.15 – Choix d'articles réellement en relation avec « *Terrorisme* » et « *Patrimoine de l'humanité* »

On peut ensuite sélectionner un certain nombre d'éléments significatifs parmi nos EN en observant celles qui sont représentées dans ces articles :

FIGURE 3.16 – Représentation des *EN person* par fréquence en nuage de mots et en contexte pour les articles sélectionnés

Ces différents éléments nous permettent de délimiter davantage les contours de nos thèmes. On constate que les EN qui nous intéressent ici sont en fait des noms de mausolées. On parle davantage d'éléments spécifiques détruits que de la notion de « patrimoine de l'humanité ». Parallèlement on ne parle pas d'actes de terroristes mais plutôt « d'islamistes radicaux », du groupe terroriste « Ansar Eddine »[3] et d'éléments relatifs à la destruction. On cherche maintenant à trouver des textes ayant la même proximité en terme de vocabulaire, pour cela on relance une requête avec le vocabulaire spécifique de nos articles :

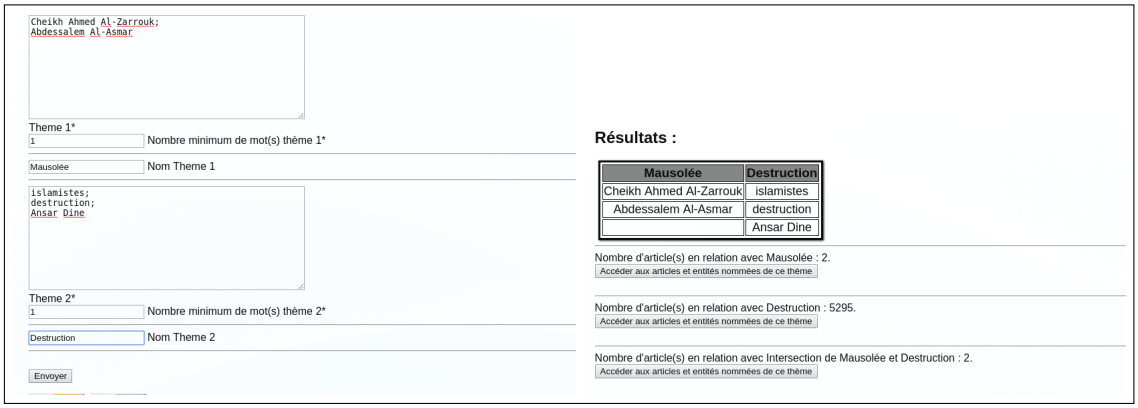


FIGURE 3.17 – Exemple de requête enrichie avec vocabulaire spécifique

Sans surprise nous obtenons en intersection les articles correspondants au vocabulaire des thèmes que nous avons définis, mais sans élément supplémentaire.

3. On trouve deux orthographes pour celui-ci *Ansar Dine* ou *Ansar Eddine*, les deux formes apparaissant dans les articles.

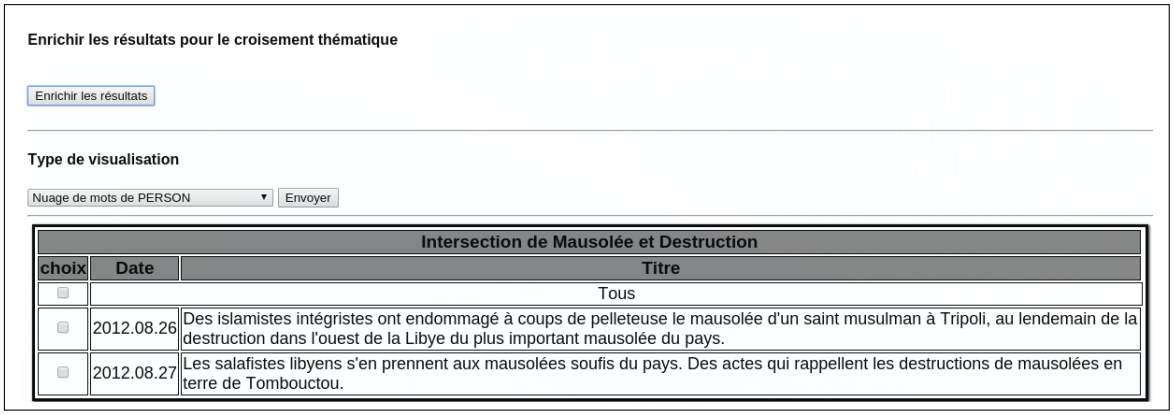


FIGURE 3.18 – Résultat de la requête utilisant un vocabulaire spécifique

Nous pouvons dans ce cas envisager un enrichissement en utilisant les coefficients de similarités [4] :

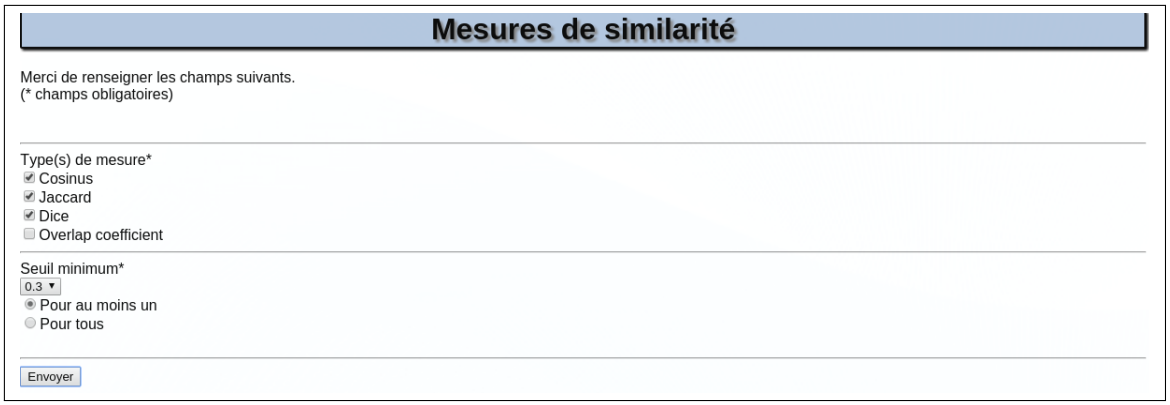


FIGURE 3.19 – Paramètre d'enrichissement thématique par coefficient de similarité

On obtient le résultat de deux articles en rapport avec la même thématique :

4. On donne ici les premiers paramètres nous ayant donné un résultat significatif.

2 résultat(s)

Mausolée et Destruction			
choix	**Date**	**Titre**	
☑	Tous		
☑	2012.08.26	Des islamistes intégristes ont endommagé à coups de pelleteuse le mausolée d'un saint musulman à Tripoli, au lendemain de la destruction dans l'ouest de la Libye du plus important mausolée du pays.	
☑	2012.08.27	Les salafistes libyens s'en prennent aux mausolées soufis du pays. Des actes qui rappellent les destructions de mausolées en terre de Tombouctou.	
Résultats par enrichissement			
☑	2012.07.01	Les islamistes d'Ansar Eddine ont commencé à détruire samedi et dimanche les mausolées de Tombouctou, menacés de subir le même sort que les Bouddhas de Bamyan, en Afghanistan.	
Cosinus :0.304452764998693	Dice : 0.304396843291996	Aucun résultat jaccard	Aucun résultat overlap
☑	2012.07.04	L'historien des religions Odon Vallet réagit aux destructions de mausolées par des islamistes à Tombouctou et revient sur le courant de pensée iconoclaste.	
Cosinus :0.300469851501681	Aucun résultat dice	Aucun résultat jaccard	Aucun résultat overlap

FIGURE 3.20 – Résultat de l'enrichissement thématique par coefficient de similarité

Nous pourrions continuer l'expérience, le procédé étant incrémentale, l'ajout de vocabulaire permettant petit à petit de définir les contours du croisement de thème, mais nous jugeons que l'expérience produit déjà des résultats significatifs que nous allons observer maintenant.

Une visualisation en réseaux de personnes nous permet de voir l'enrichissement apporté aux acteurs de ce thème avec seulement deux articles de plus que précédemment (à gauche sans enrichissement thématique par coefficient de similarité, à droite avec) :

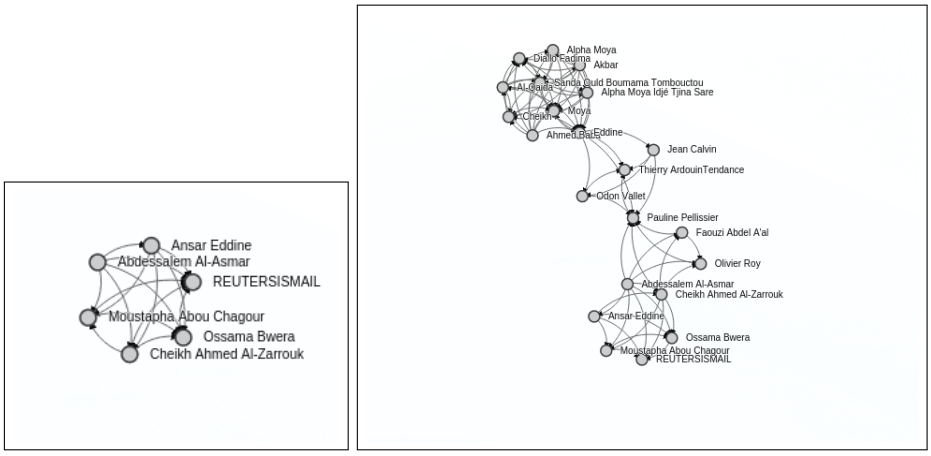


FIGURE 3.21 – Comparatif sans/avec enrichissement thématique par coefficient de similarité : Représentation en réseaux de person

On constate un mauvais typage de certaines entités nommées, comme nous le mentionnions dans le chapitre concerné. Pour autant d'autres noms de mausolée apparaissent. On peut observer la fréquence de ces mêmes entités représentées par des nuages de mots :

FIGURE 3.22 – Comparatif sans/avec enrichissement thématique par coefficient de similarité : Représentation en nuage de mots de person

La visualisation par pays est tout aussi instructive :

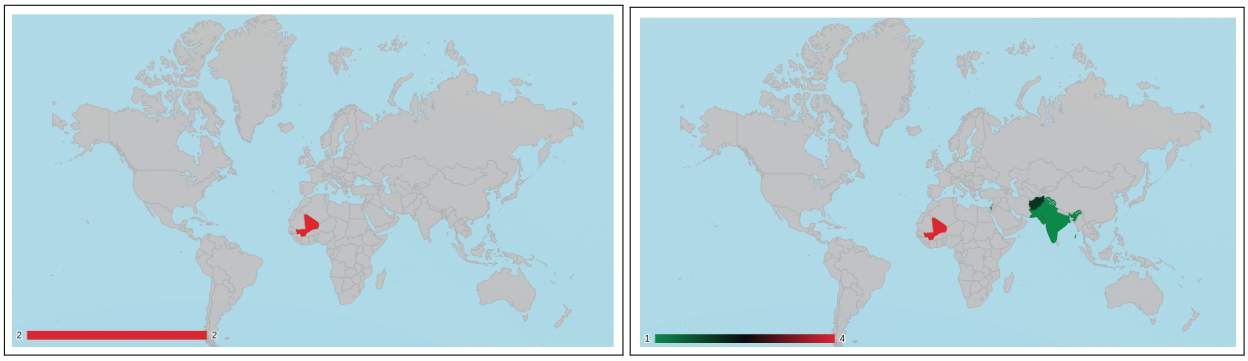


FIGURE 3.23 – Comparatif sans/avec enrichissement thématique par coefficient de similarité : Représentation par pays

On voit apparaître en plus du Mali déjà présent lors de la première requête : l'Inde, le Pakistan, l'Afghanistan et Israël.

Enfin des recherches en mode plein texte permettraient d'ajouter des éléments supplémentaires par exemple ici :


communauté. Or, les différents groupes armés qui ont pris la ville en avril (Ansar Eddine et AQMI) prônent un islam très rigoriste. Ils estiment qu'Allah seul est saint. Quand on voit les images de mausolées en miettes, on repense bien évidemment à la destruction, il y a une dizaine d'années, des bouddhas de Bâmiyân par les talibans en Afghanistan. Au Mali, ce qu'il y a de nouveau par rapport à l'épisode de 2001, c'est qu'il ne s'agit pas là de sculptures. Les islamistes s'en prennent directement à des symboles

FIGURE 3.24 – Recherche plein texte sur EN de type *date* avec enrichissement thématique par coefficient de similarité

Où l'on voit apparaître lors d'une recherche texte sur les EN de type *date*, la destruction des bouddhas de Bâmiyân en 2001 par les Talibans. Ces éléments en rapport avec cette thématique sont en-dehors du cadre temporel de notre corpus ce qui les rend d'autant plus intéressants. Cela montre aussi l'intérêt d'une recherche incrémentale, en définissant le thème recherché par étapes successives en se servant des différentes visualisations comme autant d'accès à l'information.

3.4 Conclusion

Nous avons observé dans ce chapitre l'intérêt de l'utilisation de la chaîne de traitement mise en place. Néanmoins cette chaîne est imparfaite nous observerons dans le chapitre suivant la question de l'évaluation des résultats.

Conclusion

Dans ce mémoire, nous avons tenté de répondre à la problématique :
« Comment garantir un accès à une synthèse visuelle pertinente pour un utilisateur, en partant d'un thème défini par ce dernier ? »

Pour cela nous avons d'abord standardisé notre corpus lors du chapitre 1. Nous avons ensuite rapidement défini ce que pourrait être un accès au thème du terrorisme en l'associant à une liste de mots-clefs dans la fin de ce chapitre. De la notion de thème défini par des mots-clefs, dans le chapitre 2 , nous avons ensuite posé successivement les hypothèses suivantes :

1) L'utilisateur expert était le plus adapté à définir ses propres thèmes.
2) Le croisement de ses thèmes simples pouvait donner naissance à un thème plus complexe.
3) Nous pouvions enrichir les thèmes en appliquant des mesures de similarités basées sur le vocabulaire en commun (où l'on retrouve l'idée de mots-clefs).
4) Certains éléments du thème, plus porteurs d'informations sémantiques, étaient autant d'éléments auxquels l'utilisateur pouvait souhaiter avoir accès. Nous avons choisi de représenter certains de ces éléments en nous appuyant sur la notion d'entités nommées.

Enfin le chapitre 3 nous a permis d'appliquer l'ensemble de ces constatations en réalisant l'interface web.

Plus globalement on peut voir notre travail comme l'établissement d'une chaîne de traitement dans son intégralité du corpus brut jusqu'à la visualisation des données. La construction de cette chaîne a été l'occasion de se heurter à différentes difficultés. L'aspect théorique lié à la détection de thème, le croisement thématique, l'enrichissement thématique avec l'utilisation de coefficient de similarité ou encore l'extraction d'EN en sont des exemples comme nous avons abordés de nombreuses fois. Cependant on peut aussi considérer la difficulté technique de faire communiquer les données entre différents outils, l'interfaçage de l'application web ou même la compréhension technique pour la mise en place des visualisations comme autant de plus-values.

D'autres facteurs ont conditionné la réalisation de ce travail, ainsi, aux contraintes de temps et d'expertise technique se sont ajoutées d'autres contraintes inhérentes au projet :

— Une visée applicative assumée
— Une chaîne de traitement conditionnée par des besoins de l'utilisateur
— La finalité d'avoir un prototype opérationnel en fin de stage

De ces points résulte que notre application reste imparfaite. Nous tenterons d'évaluer les éléments qui font défaut dans les sections suivantes, de mêmes que les perspectives à envisager pour garantir une amélioration. Ces remarques s'effectueront d'abord d'un point de vue général, puis suivant la chronologie présentée dans ce mémoire.

Remarques générales

On peut effectuer quelques remarques générales sur le travail réalisé :

- Textométrie : cette tâche parallèle aux développements globaux, évoquée lors du chapitre 1, qui permettait de simuler un regard d'expert et de produire une liste de mots-clefs n'est abordée qu'en surface sans réellement faire le détour de tous les outils ou techniques qui auraient pu nous aider à réaliser cette exploration du corpus.
- Détection thématique : La contrainte forte de l'utilisateur sur notre chaîne de traitement nous a poussé à assimiler la notion de thème à une liste de mots-clefs. Si l'idée n'est pas complètement étrangère à cette notion, l'observation d'autres technologies TAL en relation avec la détection de thème aurait été à envisager.

Perspectives

Ces problèmes auraient pu être résolus avec plus de temps et de recherche, parallèlement on peut envisager des améliorations de manière plus immédiate :

Type d'amélioration	Amélioration
Expressions régulières	La qualité des textes d'articles fournis est insuffisante (reste javascript, perte lors de l'extraction). On peut imaginer un traitement plus spécifique qui améliorerait les résultats.
Entités Nommées	Concernant les outils d'extraction d'entités nommées,il manque la question de l'évaluation de chacun. Par manque de temps nous n'avons pu annoter et tester chacun des outils avec les habituels rappel, précision et f-mesure. L'annotation et le test de ces standards seraient ici à prévoir.
Entités Nommées	L'abandon de certains outils par manque de temps devrait être revu, de façon à obtenir plus de matière pour les EN extraites. Si nous proposons certaines solutions (comme sur mXS ou sur DARK) pour améliorer ou produire des résultats, nous n'avons pu les mettre en pratique. Ces solutions peu coûteuses à mettre en place en terme de temps sont largement envisageables.
Entités Nommées	Nous n'avons que peu traité le croisement d'entités nommées issues de l'extraction de différents outils. De même un croisement avec des connaissances extérieures a été envisagé mais non implémenté. Cette implémentation aurait permis de filtrer les EN non pertinentes. Cela aurait aussi permit d'obtenir des informations supplémentaires basées sur le web sémantique. Ces solutions sont a étudier.
Enrichissement par similarité	Quelque peu conditionné par l'architecture matérielle, nous ne proposons l'enrichissement par similarité que dans les cas d'intersection thématique, considérant que les thèmes de base sont en général très fournis. De plus on ne propose pas à l'heure actuelle un enrichissement possible sur une sélection d'articles. Il conviendrait de rendre ces options accessibles à l'utilisateur.
Enrichissement par similarité	Le choix de considérer un thème comme un seul article est plus que discutable, il apparaît que l'on devrait ici plutôt chercher la similarité en nous basant sur des couples *article+article* plutôt que sur un couple *thème+article*. Cependant comme nous l'avons vu cette solution prend du temps et de l'espace (plusieurs dizaines de Go sont à prévoir)
Visualisation	On peut déjà envisager des améliorations concernant les visualisations. Par exemple l'usage de graphiques pour indiquer la fréquence associés aux nuages de mots ou encore un calendrier interactif permettant de voir sur quelles dates sont concentrés les événements.
Visualisation	Certaines visualisations comme les réseaux mériteraient d'être réimplémentés pour nos besoins. Par exemple la possibilité en cliquant, de déconnecter certains réseaux ou nœuds de façon à alléger une visualisation saturée d'informations.
Visualisation	Une avancée intéressante aurait été de traiter la question des patrons syntaxiques. En effet nous nous sommes arrêtés à l'échelle de l'article, or si nous avions pu descendre jusqu'à celle de la phrase et établir des prédicats entre entités nommées nous aurions pu déterminer par exemple l'appartenance d'une entité *person* à une entité *organization*. On aurait pu aussi établir un lien entre EN de manière plus précise. Par exemple en se basant sur les relations syntaxiques et la valence des verbes on peut imaginer produire des représentations comme : $Taliban \rightarrow ont\ détruit \rightarrow Bouddhas\ de\ Bâmiyân$

TABLE 3.1 – Perspectives

A toutes ces remarques on peut adjoindre une remarque générale dont nous n'avons pu trouver la solution. En effet, si nous avons montré l'utilité de l'outil sur certaines requêtes, se pose la question de l'évaluation des visualisations. Comment quantifier la valeur de ces informations ? Cette question reste en suspens et devrait être traitée en priorité dans le cadre d'un approfondissement sur les représentations visuelles.

BIBLIOGRAPHIE

[Azpeitia et al., 2014] Azpeitia, A., Cuadros, M., Gaines, S., and Rigau, G. (2014). Nerc-fr: Supervised named entity recognition for french. In *Text, Speech and Dialogue*, pages 158–165. Springer. – Cité pages 40, 41 et 42.

[Bittar, 2010] Bittar, A. (2010). *Building a TimeBank for French: a reference corpus annotated according to the ISO-TimeML standard*. PhD thesis, Paris 7. – Cité page 33.

[Bittar et al., 2011] Bittar, A., Amsili, P., and Denis, P. (2011). French timebank: un corpus de référence sur la temporalité en français. In *TALN 2011-Traitement Automatique des Langues Naturelles*, volume 1, pages 259–270. Laboratoire d'Informatique de Robotique et de Microélectronique. – Cité page 33.

[Cox and Ellsworth, 1997] Cox, M. and Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. *Proceedings of visualisation 97, phoenix*, pages 235–244. – Cité page 9.

[Fleury, 2007] Fleury, S. (2007). Le trameur, manuel d'utilisation. – Cité page 22.

[Gravier et al., 2012] Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., and Galibert, O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In *LREC-Eighth international conference on Language Resources and Evaluation*, page na. – Cité page 37.

[Hamon et al., 2014] Hamon, T., Pleplé, Q., Paroubek, P., Zweigenbaum, P., and Grouin, C. (2014). Analyse automatique de textes littéraires et scientifiques: présentation et résultats du défi fouille de texte deft2014. *Actes du dixième DÉfi Fouille de Textes*, page 3. – Cité page 28.

[Lebart and Salem, 1994] Lebart, L. and Salem, A. (1994). Statistique textuelle. *Paris: Dunod,| c1994*, 1. – Cité pages 21 et 22.

[Longo and Todirascu, 2010] Longo, L. and Todirascu, A. (2010). Une étude de corpus pour la détection automatique de thèmes. *Proceedings of the 6th journées de linguistique de corpus*. – Cité page 26.

[Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press. – Cité page 28.

[Nouvel, 2012] Nouvel, D. (2012). *Reconnaissance des entités nommées par exploration de règles d'annotation*. PhD thesis. – Cité pages 32, 37, 38 et 42.

[Omrane et al., 2011] Omrane, N., Nazarenko, A., and Szulman, S. (2011). Les entités nommées: des clés linguistiques pour la conceptualisation. In *22èmes journées francophones d'ingénierie des connaissances*, pages 435–450. – Cité page 44.

[Poibeau, 2011] Poibeau, T. (2011). *Traitement automatique du contenu textuel*. Lavoisier, Paris. – Cité page 32.

[Rosset et al., 2011] Rosset, S., Grouin, C., and Zweigenbaum, P. (2011). *Entités nommées structurées: guide d'annotation Quaero*. LIMSI-Centre national de la recherche scientifique. – Cité pages 33 et 42.

[Ruiz and Poibeau, 2015] Ruiz, P. and Poibeau, T. (2015). Combining open source annotators for entity linking through weighted voting. In *Proceedings of* SEM 2015. Fourth Joint Conference on Lexical and Computational Semantics*. – Cité page 47.

[Salem et al., 2003] Salem, A., Lamalle, C., Martinez, W., Fleury, S., Fracchiolla, B., Kuncova, A., and Maisondieu, A. (2003). Lexico3–outils de statistique textuelle. manuel d'utilisation. *Syled-CLA2T, Université de la Sorbonne nouvelle–Paris*, 3. – Cité pages 21 et 22.

[Schwer and Tovena, 2009] Schwer, S. and Tovena, L. (2009). Ontologies temporelles et sémantique de la temporalité. *XVIèmes rencontres de Rochebrune, Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels*. – Cité page 33.

[Yau, 2013] Yau, N. (2013). *Data Visualisation*. Édition Eyrolles, Paris. – Cité page 56.

ANNEXE A

Liste thématique terroriste

La liste présentée ici est issue du travail effectué sur les logiciels le Trameur et Lexico3, dans la section 1.5 p.22. Elle nous a servi pour simuler le regarde de l'expert et caractériser le thème « Terrorisme ». On trouvera ici les liens présentant les outils :

— Lexico3 : http://www.tal.univ-paris3.fr/lexico/
— Trameur : http://www.tal.univ-paris3.fr/trameur/

1. Abou Khaled Al-Souri
2. Abou Mohammad Al-Joulani
3. AQMI
4. Al-Qaida
5. Al-Qaïda
6. Al-Souri
7. Amedy Coulibaly
8. assassinat barbare
9. attentat
10. Attentat
11. Ayman Al-Zawahiri
12. Ben Laden
13. Boko Haram
14. brigade de femmes Al-Khansaa
15. califat
16. Califat
17. califette
18. califettes
19. Califettes
20. Charlie Hebdo
21. Cherif et Said Kouachi
22. Cherif Kouachi
23. Daech
24. Djamel Beghal
25. djihad
26. Djihad
27. djihadiste
28. Djihadiste
29. djihadistes
30. Djihadistes
31. EI
32. enlèvement
33. Etat islamique
34. Etat Islamique
35. État Islamique
36. frères Kouachi
37. Frères Kouachi
38. Front Al-Nosra
39. gilets pare-balles
40. Grand Satan américain
41. Guantanamo
42. Hayat Boumeddienne
43. je suis Charlie
44. Je Suis Charlie
45. jihad
46. Jihad
47. kalachnikov
48. Kalachnikov
49. kamikaze
50. Kamikaze
51. kamikazes
52. Kamikazes
53. Kenji Goto
54. magazine Al-Shamikha
55. marche républicaine
56. massacre
57. Mohammed Merah
58. Mohamed Merah
59. Merah
60. Tsarnaïev
61. Tamerlan Tsarnaïev
62. Dzhokhar Tsarnaïev
63. Djokhar Tsarnaïev
64. Frères Tsarnaïev
65. Tamerlan Tsarnaev
66. Dzhokhar Tsarnaev
67. Djokhar Tsarnaev
68. Tsarnaev
69. Tamerlan Tsarnaïev
70. Djokhar Tsarnaïev
71. organisations terroristes
72. organisation terroriste
73. réseaux sociaux
74. Saïd Kouachi
75. sécurité
76. services de renseignement
77. terrorisme
78. Terrorisme
79. terroriste
80. Terroriste
81. tuerie
82. vidéos de décapitation

ANNEXE B

UNIFORMISATION

L'uniformisation expliquée dans la section 1.3 p.14 est en grande partie réalisée par la boucle suivante :

```
1  foreach my $cle (sort{$a<=>$b}(keys %articles))
2  {
3    my $ligne = $articles{$cle};
4    my $titre = checkPatron($ligne,'<filname="SURF.*?>(.*?)<');
5    $titre = supprimeMeta($titre);
6    $titre = supEsp($titre);
7    my $date = checkPatron($ligne,'<AAMMJJ="(.*?)">');
8    $date =~ s/(....)(..)(..)/$1.$2.$3/;
9    my $article =checkPatron($ligne,'<filname="PROF.*?>(.*)');
10   $article = supprimeMeta($article);
11   $article = supEsp($article);
12   if(($titre ne "")and($article ne ""))
13   {
14     #imprimeEcran($titre,$date,$article);
15     imprimeFile($titre,$date,$article,$nameOut);
16   }
17 }
```

Une fois les fichiers sous des formats uniformes et stockés dans une table de hash à raison de un article sur une ligne par clé de cette table (on a, pour cela, utilisé une première expression régulière pour délimiter un bloc de texte brut correspondant à un article). On parcours ensuite la table en suivant l'ordre des clé (ligne 1), où chacune des clés est un numéro entre 1 et le nombre maximale d'articles. On stocke dans la variable *$ligne* le fichier sur une ligne (ligne 3), toutes les opérations suivantes seront effectuées sur cette ligne. La fonction *checkPatron()* prend deux arguments

- Une ligne de texte
- Un patron dont un élément entre parenthèses que sera renvoyé comme résultat

Elle est appelée en lignes 4, 7, et 9 et permet d'extraire respectivement le titre, la date et le corps de l'article. Les fonctions *supprimeMeta()* et *supEsp()* permettent respectivement un nettoyage de caractères et d'espaces elles seront appelées sur le texte du titre et de l'article (lignes 5, 6, 10, 11). On effectue aussi un reformatage de la date. Après tout cela on imprime l'extraction uniquement si les textes du titre et de l'article ne sont pas vide (condition ligne 12 et impression ligne 15).

ANNEXE C

CROISEMENT THÉMATIQUE

La recherche thématique abordée dans les sections 2.2.2 p.26 et 2.4.2 p.50 s'effectue par une transformation des articles en objet. On trouvera ci-dessous e, exemple, l'objet Article en Perl.

```
# Traitement de l article
package Article;

# classes d un objet Article
sub new
{
  my ($class,$num,$dateArt,$titre,$article) = @_;
  my $this =
  {
    "num" => $num,
    "dateArt" => $dateArt,
    "titre" => $titre,
    "article" => $article

  };
  bless($this, $class);
  return $this;
}
```

On vérifie ensuite la présence des mots clés dans l'objet selon le critère seuil de l'utilisateur :

```
# récupération des indices si le thème est présent
sub checkCorpus
{
  my $refLexique1 = shift;
  my $refLexique2 = shift;
  my $refArticle = shift;
  my $refHashIndice1 = shift;
  my $refHashIndice2 = shift;
  my $nbrMot1 = shift;
  my $nbrMot2 = shift;
  my $compteurArticle=0;
  foreach my $el(@$refArticle)
  {
```

```
14      # pour chaque article
15      my @compteMot1=();
16      my @compteMot2=();
17      # on récupère le texte (titre+article)
18      my $texte = "$el->{'titre'} $el->{'article'}";
19      $compteurArticle++;
20      # on consulte les mots du thème pour savoir si ils sont présents
            dans le texte si oui, on stocke
21      foreach my $cle (sort(keys %$refLexique1))
22      {
23        if($texte=~ m/( $cle.?)/)
24        {
25          push(@compteMot1,$cle);
26        }
27      }
28      foreach my $cle (sort(keys %$refLexique2))
29      {
30        if($texte=~ m/( $cle.?)/)
31        {
32          push(@compteMot2,$cle);
33        }
34      }
35      # si le nombre de mot correspond à celui attendu par l utilisateur
            alors le texte est considéré en accord avec le thème
36      if(scalar(@compteMot1) >= $nbrMot1)
37      {
38        $$refHashIndice1{$compteurArticle}=$el->{'num'};
39      }
40      if(scalar(@compteMot2) >= $nbrMot2)
41      {
42        $$refHashIndice2{$compteurArticle}=$el->{'num'};
43      }
44    }
45  }
```

Ce code assez basique permet un parcours de tous les objets articles (ligne 12) et de vérifier si les mots souhaités sont présents. Pour cela on récupère le texte de l'objet (ligne 18), on vérifie pour chacun des mots des deux lexiques (lignes 21 et 28) si les mots sont présents (lignes 23 et 30) , si oui on les stocke dans des tableaux. On vérifie ensuite le nombre de mots présents dans l'article (lignes 36 et 40) si oui, on récupère l'Id de l'article.

ANNEXE D

BASE DE DONNÉES

La consultation de la base de données abordée en section 2.3.5 p.48 et en 2.4.2 p.53 s'effectue grâce à une série de module Perl dont on peut trouver la référence ci-dessous :

— Pour *DBI* :
http://dbi.perl.org/

— Pour *DBD :: SQLite* :
http://search.cpan.org/~msergeant/DBD-SQLite-0.31/lib/DBD/SQLite.pm

La plupart des accès en lecture de la base de données s'effectuant ensuite grâce à des fonctions comme celle présentée ci-dessous :

```
1  sub consultBdd
2  {
3    my $bdd= shift;
4    my $refIndice = shift;
5    my $table = shift;
6    my $en = shift;
7    my $dbh = DBI->connect
8    (
9      "dbi:SQLite:dbname=$bdd","","",
10     {
11       RaiseError     => 1,
12       sqlite_unicode => 1,
13     }
14   );
15   my %results;
16   my $requete = $dbh->prepare("SELECT $en FROM $table WHERE (id =
       ?);");
17   foreach my $id(@$refIndice)
18   {
19     $requete->execute($id);
20     my @temp = $requete->fetchrow_array();
21     my @parseTemp;
22     foreach my $el(@temp)
23     {
24       push(@parseTemp,split(/;/,$el));
25     }
26     push(@{$results{$id}},@parseTemp);
27   }
28   $requete->finish;
29   # Déconnection de la base de données
```

```
30    $dbh->disconnect();
31    return(\%results);
32  }
```

Après avoir récupéré :

— le chemin vers la base de données (ligne 3)
— la référence vers un tableau contenant les id d'articles que l'on souhaite consulter (ligne 4)
— le nom de la table que l'on va interroger (ligne 5)
— le nom du champ de la table dont on veut obtenir le contenu (ligne 6)

On se connecte à la BDD (ligne 7) et l'on initialise une table de hash qui contiendra les résultats (ligne 16), après quoi on prépare la requête vers la table (ligne 17). Dans une boucle qui va traiter tous les id d'articles (ligne 18) on va successivement :

— exécuter la requête (ligne 20)
— récupérer les résultats dans un tableau (ligne 21)
— initialiser un tableau servant à récupérer les résultats « parsés » du tableau de la ligne précédente (ligne 22)
— pour chaque résultat correspondant à la requête, on découpe (parse) la ligne en s'appuyant sur le caractère « ; » et on stocke le résultat dans le tableau prévu(ligne 24) [1]
— On copie (ligne 26) sous forme de référence dans la table de hash *%results* à la clé *$id* le tableau que l'on vient de remplir à la ligne précédente.

Enfin lignes 28, 30, 31 on termine la requête, on se déconnecte de la base de données et on renvoi comme résultat de notre fonction la table résultat sous forme de référence.

1. on rappelle que nos entités nommées dans la base de données sont de la forme `Christian Charrière-Bournazel;Dray;Jean-Claude Marin;Jean-Paul Huchon;Julien Dray;Léon Lev Forster;Nicolas Sarkozy;Pascale Robert-Diard`

ANNEXE E

COEFFICIENT DE SIMILARITÉ

On traite ici du de la question de l'enrichissement thématique abordé dans les sections 2.2.3 p.28 et 2.4.2 p.53. La partie la plus importante du code est présentée ici avant d'être commentée :

```
1  sub calculCoeff
2  {
3    my $refList1 = shift;
4    my $refList2 = shift;
5    # choix des mesures recherchées
6    my $getCosinus = shift;
7    my $getDice = shift;
8    my $getJaccard = shift;
9    my $getOl = shift;
10   # choix du seuil et de la cible (tous au dessus du seuil ou
         seulement un élément)
11   my $seuil = shift;
12   my $cible = shift;
13
14   # nbr de mot dictionnaire 1 et 2
15   my $card1 = scalar @{$refList1};
16   my $card2 = scalar @{$refList2};
17   # Intersection des deux documents (matching coefficient)
18   # les mots qui apparaissent dans les deux documents
19   my $lc = List::Compare->new('-u', '-a',$refList1, $refList2);
20     my $intersection = scalar($lc->get_intersection);
21
22   # mesures de similarité  :
23
24   my %table;
25   if($getCosinus)
26   {
27     my $cosinus =  $intersection / sqrt($card1 * $card2);
28     if($cosinus >= $seuil)
29     {
30       $table{'cosinus'} =$cosinus;
31     }
32   }
33   if($getDice)
34   {
35     my $dice = 2 * $intersection / ($card1+$card2);
36     if($dice >= $seuil)
```

```
37      {
38        $table{'dice'} =$dice;
39      }
40    }
41    if($getJaccard)
42    {
43      # Union des deux documents
44      my $union = $card1 + $card2 - $intersection;
45      my $jaccard = $intersection / $union;
46      if($jaccard >= $seuil)
47      {
48        $table{'jaccard'} = $jaccard;
49      }
50    }
51    if($getOl)
52    {
53      my $ol = $intersection / min($card1,$card2);
54      if($ol >= $seuil)
55      {
56        $table{'ol'} = $ol;
57      }
58    }
59    my $compteur=0;
60    foreach my $cle(keys %table)
61    {
62      $compteur++;
63    }
64    if(($compteur>0)and($compteur>=$cible))
65    {
66      return(\%table);
67    }
68  }
69
70  # fonction minimum pour overlap coefficient
71  sub min
72  {
73    my ($cardx,$cardy) = @_;
74    if($cardx > $cardy)
75    {
76      return $cardy;
77    }
78    else
79    {
80      return $cardx;
81    }
82  }
```

On commence par récupérer les différentes options de la ligne 3 à la ligne 12 : les listes pour lesquelles on va effectuer le comparatif avec les coefficients de similarité, le type des mesures recherché (1 on recherche, 0 non), le seuil ([0.1-0.9] avec un pas de 0.1) et la façon de gérer ce seuil (toutes les mesures doivent être supérieure ou une seule). On récupère ensuite le cardinal de des éléments des listes, soit le nombre de mots (ligne 15 et 16). On récupère le le nombre de vocabulaire en commun (ligne 19 et 20), la simplification selon l'idée que le terme est présent ou non revient à considérer un vecteur booléen dont la position des mots est ici sans importance. On applique

ensuite les différentes formules sur la base de ces intersections. On récupère les résultats en ayant vérifier pour chaque cas si on souhaite la mesure (avec les variables $get*) et si ce résultat est supérieur au seuil fixé par l'utilisateur. On renvoit ensuite une référence vers une table de hash résultat si le nombre de résultat correspond à l'attente de l'utilisateur. A remarquer aussi la fonction min qui renvoi le minimum entre le cardinal de x et celui de y.

ANNEXE F

PERL DANCER 2

Nous abordons le module Dancer 2 dans le cadre de la construction de notre interface web à la section 2.4.2 p.50. On peut retrouver ici les références citées :

— Le site de Dancer :
http://perldancer.org/
— Sur CPAN :
http://search.cpan.org/ xsawyerx/Dancer2-0.163000/lib/Dancer2.pm
— Sur GitHub :
https://github.com/PerlDancer/Dancer2
— Ou encore le site wikipédia :
https://en.wikipedia.org/wiki/Dancer_(software)

Nous détaillons ci-après le processus d'implémentation de l'interface. Dancer est très simple à mettre en place une fois installé (*sudo cpan install Dancer2*), on crée l'application avec :

```
dancer2 -a nomApplication
```

Cette application simulera ensuite un serveur qu'on peut lancer via :

```
plackup -r bin/app.psgi
```

On peut ensuite consulter les pages sur le port 5000 en local :

FIGURE F.1 – Appel de l'interface via Dancer

Il ne reste ensuite qu'à définir des « chemins » en fonction des actions à effectuer comme ici :

```
any '/Intersection' => sub
{
  $titre = "Intersection Thématique";
  my %param = params;
  my @theme=split(/ /,$param{'id'});
  my $nom=$param{'nom'};
  my $result;
  if(0==scalar(@theme))
  {
    $result.="Erreur, le croisement thématique n'est pas effectué,
       merci de retourner en page d'accueil<br/>";
  }
  else
  {
    $result.=tableauHtmlXcol(\@theme,$nom);
  }
  template 'Intersection',{titre =>"$titre $nom",mot=>"$result"};
};
```

Où l'on récupère le résultat d'une requête *get* ou *post*, pour afficher la page *Intersection*. La récupération des paramètres de la requête s'effectue via la table de hash *%param*. On effectue ensuite nos traitements avant de renvoyer un résultat sous forme de chaîne de caractères dans *$result*. Ce texte est ensuite incorporé au template global de notre interface web. Le même fonctionnement se reproduit quelque soit la requête. On peut ajouter qu'il est très facile de définir les templates du site en HTML/CSS comme ici :

```
<!DOCTYPE html PUBLIC "-//W3C XHTML 1.0 Transitional//EN"
   "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<!-- balise minimale pour l\'existance du doc -->
<html xmlns="http://www.w3.org/1999/xhtml">
  <!-- informations sur le document -->
  <head>
    <link rel="stylesheet" type="text/css" href="css/styles.css"/>
    <!-- titre -->
    <title><% titre %></title>
    <!-- ensemble des balises meta qui sont informatives sur la
       structure du document -->
    <meta http-equiv="Content-Type" content="text/html;
       charset=UTF-8"/>
    <!-- l\'auteur -->
    <meta name="author" content="Ferguth Johan"/>
    <!-- email -->
    <meta name="reply-to" content="jferguth - @ - gmail.com"/>
    <!-- description de la page-->
    <meta name="description" content="Extraction d'informations,
       Entités nommées, Croisement thématique, Visualisation
       thématique, Base de données, lattice, inalco, master 2, tal"/>
    <!-- mots-clés -->
    <meta name="keywords" content=""/>
    <!-- favicon -->
    <link rel="icon" href="images/favicon.ico" />
  <!-- fermeture des informations -->
```

```
22    <% entete %>
23  <script type="text/javascript" src="javascripts/jquery.js"></script>
24    </head>
```

Ce code correspond à l'entête des pages de notre application. Le seul point intéressant à expliquer ici est que l'on n'écrit le code qu'une fois et qu'on peut le changer au moment de l'envoi d'information via le template, en utilisant des variables comme dans le ligne huit du code ($< \%titre\% >$) ou encore à la ligne 22 ($< \%entete\% >$). Ces éléments nous permettent par exemple ici d'ajouter un titre différent à chaque page ou encore d'ajouter du code javascript dans nos pages web.

Le passage d'informations s'effectuant par requête *get* ou *post*, on utilisera la plupart du temps des formulaires HTML pour interagir avec l'utilisateur.

ANNEXE G

CORTEXT

Nous abordons les visualisations traitées dans Cortext dans le chapitre 2.4.3 p.56 à titre d'exemple comme outil de TAL produisant des visualisations. On pourra consulter la page de Cortext pour plus de détail : http://www.cortext.net/. En effectuant quelques tests avec Cortext sur notre corpus on obtient les visualisations suivantes, à titre indicatif, où on observera la différence avec notre approche, basé sur les EN :

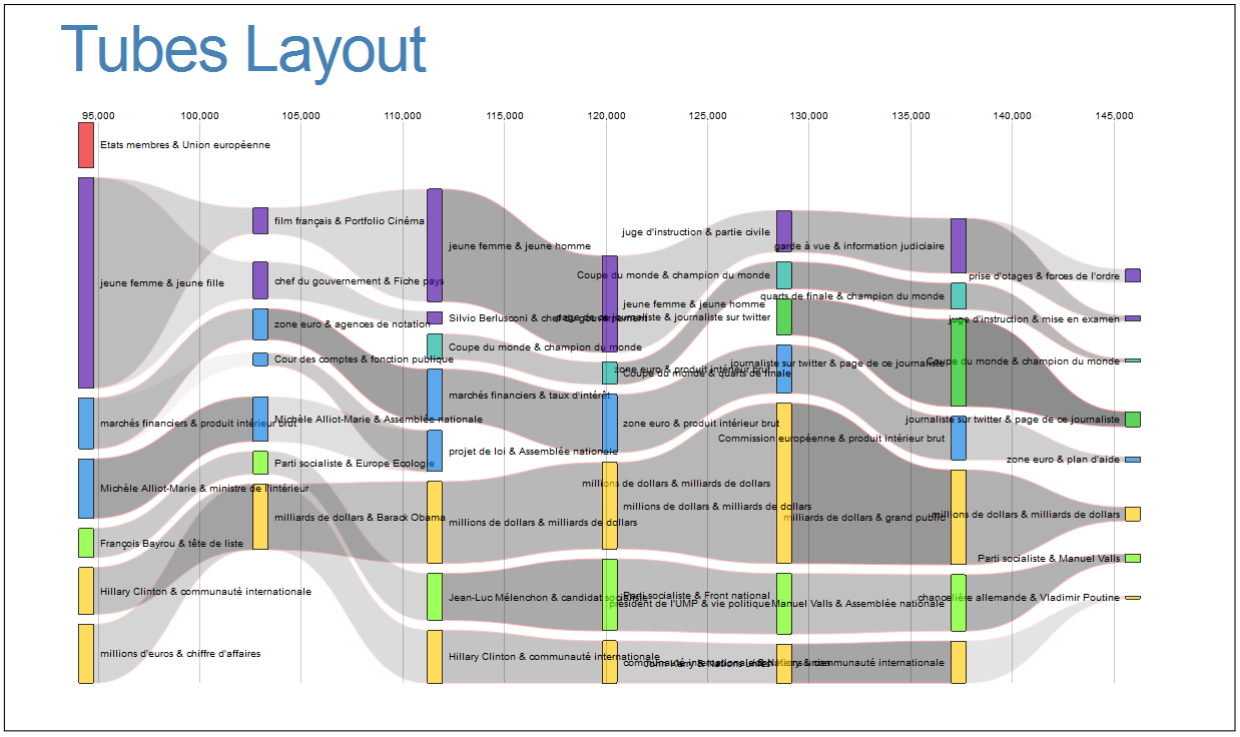


FIGURE G.1 – Évolution des termes en fonction du temps

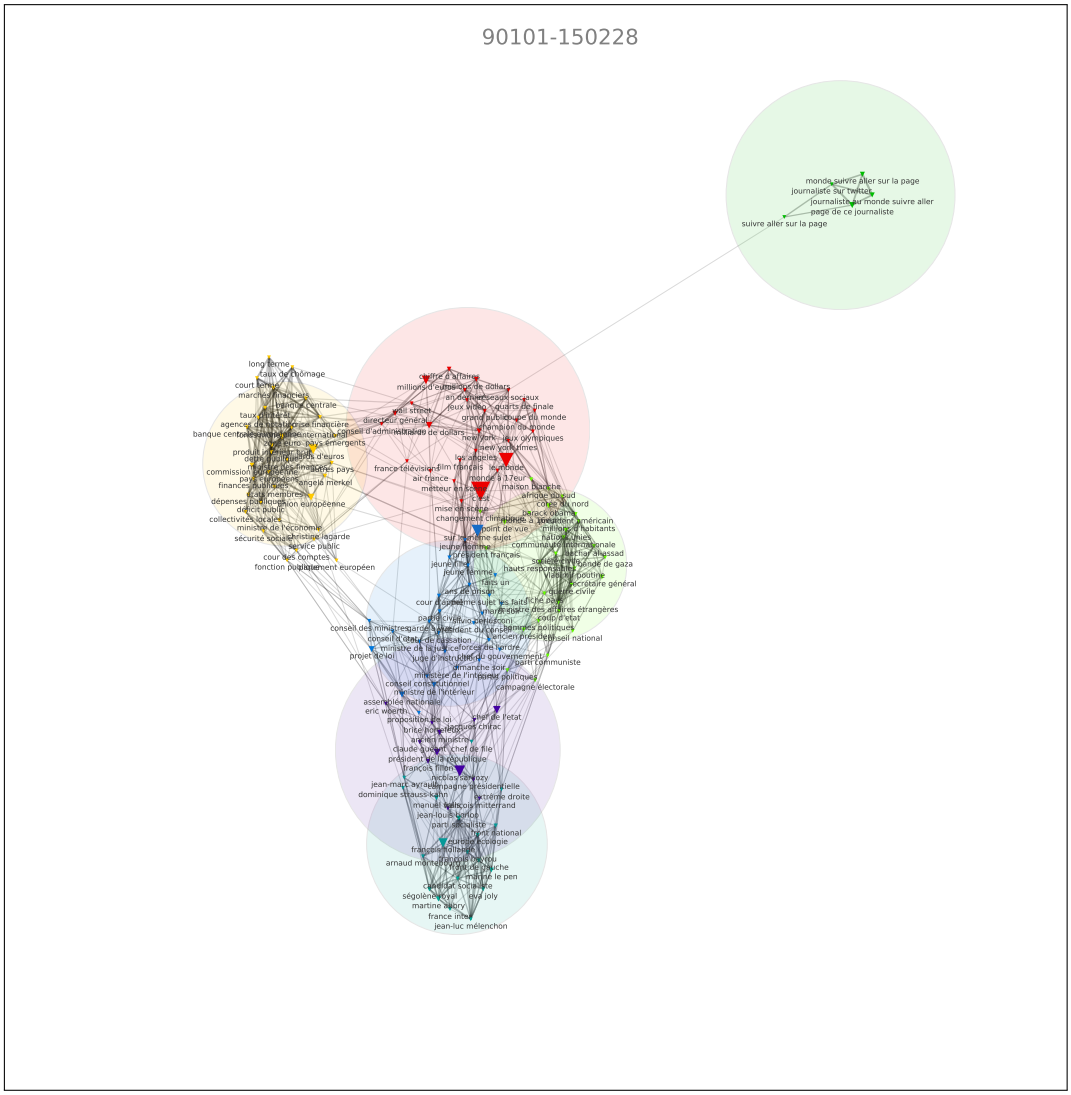


FIGURE G.2 – Clusters de termes les plus présents sur le corpus

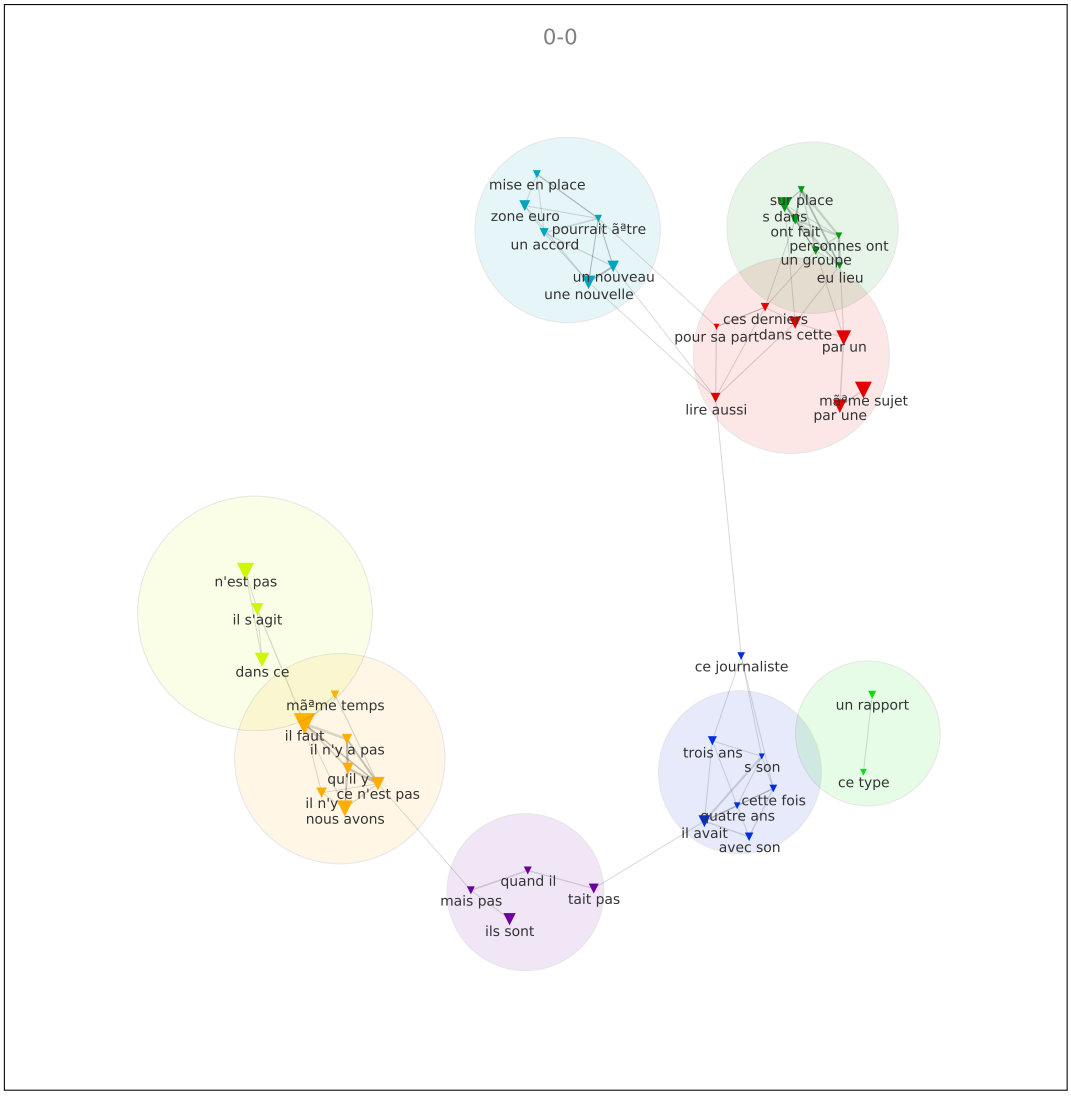


FIGURE G.3 – Clusters de termes les plus présents sur le corpus, réduction du nombre d'informations demandées

Index