Institut National des Langues et Civilisations Orientales

M1 Traitement Automatique des Langues

Projet réalisé dans le cadre du cours
"Extraction d'informations"

Modélisation d'une ontologie du terrorisme et du djihadisme sur base des articles de la rubrique "International" du journal "Le Monde" de 2010 à 2016

Sotiria Bampatzani
Morgane Dehareng

Supervisé par:
Mme Huguette Rigot

24 mai 2017

Table des matières

Introduction

Notre projet consiste en une tentative de modélisation d'une ontologie du terrorisme et du djihadisme à partir d'articles de la rubrique "International" du journal "Le Monde" de 2010 à 2016. Cette problématique au coeur de l'actualité fait couler beaucoup d'encre et il nous est paru intéressant d'ajouter à ce tableau le point de vue d'un linguiste.

Initialement, notre recherche portait sur la notion de terreur. Prenant petit à petit conscience de l'étendue du sujet et de l'absence d'ontologies sur le sujet, nous avons dévié vers le terrorisme, passant ainsi d'un concept abstrait à une triste réalité. Cette nouvelle thématique était plus délicate que la précédente, ce qui explique ici aussi le manque de sources concernant une éventuelle ontologie déjà existante. Notons pourtant le livre de M. Thierry Balzacq, "Théories de la sécurité. Les approches critiques" [1], malheureusement disponible uniquement pour consultation et sous réserve à la Faculté de Droit.

La méthodologie que nous utiliserons est de type bottom-up : nous commencerons par constituer et étiqueter le corpus, puis nous procéderons à l'extraction d'entités nommées sur lesquelles nous pourrons par la suite nous baser pour modéliser notre ontologie.

Modélisation d'une ontologie du terrorisme et du djihadisme

Constitution du corpus

Nous voulions au départ nous concentrer sur la notion de terreur dans les articles du journal "Le Monde" de 2010 à 2016, toutes rubriques confondues. Nous pensions pouvoir extraire automatiquement tous les articles concernés à l'aide d'un script Perl. Ce script s'est toutefois révélé inefficace. En effet, le résultat de l'extraction n'affichait que le titre de l'article et pas son contenu. De plus, il fallait aussi distinguer le contenu payant du contenu gratuit. Nous avons donc opté pour une extraction manuelle des articles.

Nous avons trié les articles par année et par rubriques : à la une, culture, disparitions, économie, idées, international, planète, politique, société, sport, technologies et voyage. Nous pensions analyser notre corpus à l'aide de l'outil textométrique Lexico3 et nous avons donc procédé au balisage, dont voici un exemple :

<annee=2010>

<rubrique=societe2010>

<article=01>

Nous avons ensuite utilisé la commande "cat" pour concaténer tous les fichiers et ainsi obtenir le corpus complet (7 820 418 occurrences). En vue d'une analyse avec Lexico3, nous avons également prévu deux versions du corpus : une encodée en UTF-8 et une encodée en Windows-1252 (ANSI). Notre arborescence de travail ressemblait à celle ci-dessous :

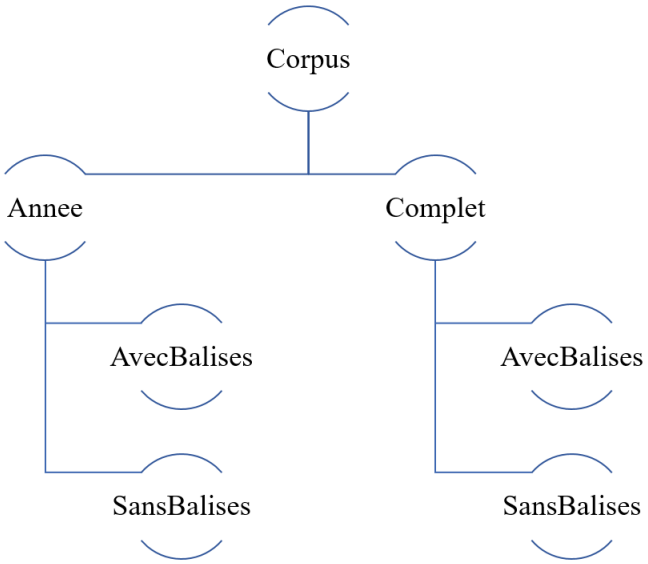

FIGURE *1 – Arborescence 1*

Au vu de l'ampleur de notre corpus, nous pensions qu'il était plus pertinent de travailler sur une seule rubrique. Nous avons constaté que la rubrique "international" était celle qui regroupait le plus d'informations et nous avons fait le choix de nous concentrer uniquement sur celle-ci. Après quelques recherches, nous nous sommes aperçues que la notion de terreur était très souvent liée à la notion de terrorisme dans notre corpus et nous avons restreint notre sujet au terrorisme.

Ce changement impliquait de pouvoir isoler les différents groupes terroristes dans notre corpus, et donc de procéder à une reconnaissance des entités nommées. Lexico3 ne nous offrant pas cette possibilité, nous avons opté pour un étiquetage du corpus à l'aide de SEM, un logiciel intégrant la reconnaissance des entités nommées.

Étiquetage du corpus

Pour réaliser l'étiquetage, nous avons utilisé le logiciel SEM, un segmenteur-étiqueteur du français développé par le laboratoire LaTTiCe. Ce logiciel dispose de deux versions : une version en ligne et une version en console. Cette dernière, optimisée pour un environnement Unix, est assez complexe à prendre en main, raison pour laquelle nous avons préféré la version en ligne.

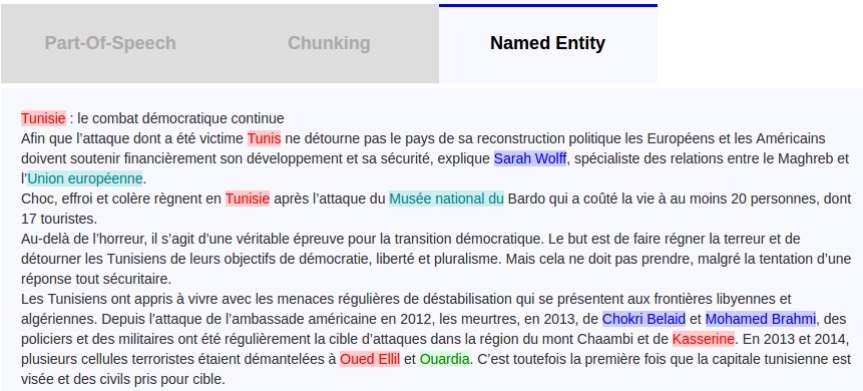

FIGURE *2 – L'interface graphique de l'étiqueteur SEM*

Nous ne pouvions évidemment pas étiqueter l'entièreté du corpus d'un coup. Nous avions pensé qu'il serait plus aisé de procéder par année, mais nos sous-corpus étaient tout de même trop volumineux pour le site. Nous avons donc morcelé le corpus en plusieurs parties pour en faciliter l'étiquetage, comme montré ci-dessous :

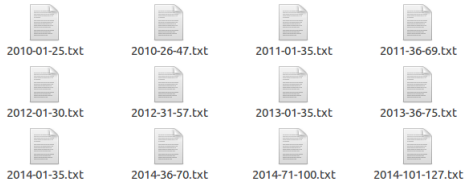

FIGURE *3 – Sous-corpus*

Nous avons supprimé les balises, devenues inutiles, mais nous avons conservé une trace du corpus balisé. Notre arborescence se décomposait donc ainsi :

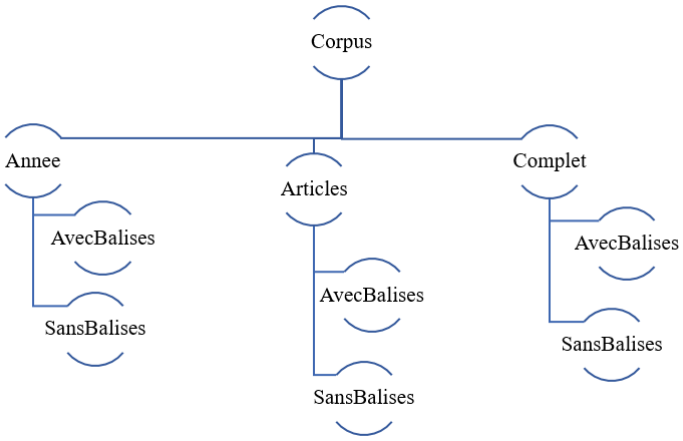

FIGURE *4 – Arborescence 2*

Nous avons repris la structure de cette arborescence pour créer un répertoire qui contiendrait tous les fichiers étiquetés. Le résultat de l'étiquetage avec SEM en ligne peut générer soit un fichier au format .conll soit un fichier au format .text. Ne sachant pas encore comment nous allions exploiter nos données, nous avons téléchargé les deux formats pour chaque étiquetage réalisé. Comme précédemment, nous avons ensuite concaténé les fichiers pour recréer les sous-corpus par année et le corpus complet.

Recherche des entités nommées

Pour la recherche des entités nommées, nous avons écrit un script Perl qui extrait ces entités sur base d'un fichier .text. Nous nous sommes pour cela servies des expressions régulières pour reconnaitre les différentes catégories d'entités nommées (personnes, lieux, compagnies, organisations) et pour réaliser des traitements sur les données propres à chaque catégorie. Le résultat de cette extraction est un fichier au format .txt contenant la liste des entités extraites.

Il ne faut pas oublier qu'une ontologie repose sur le principe sujet-relation-prédicat. Nous disposions de listes de sujets et de prédicats potentiels, mais nous n'avions sous la main aucune relation. Nous avons alors extrait plusieurs syntagmes verbaux qui correspondaient au verbe avoir ou être à l'infinitif suivi d'un participe passé. Ces syntagmes nous permettaient de nous faire une idée plus précise des relations possibles au sein du corpus (par exemple : avoir tué, avoir arrêté, être attaqué). Tous les termes trouvés ont ensuite été classés par fréquence à l'aide de la commande :

cut -f1 fichier.txt | sort | uniq -c | sort -rg > fichier-freq.txt

Nous avons réalisé l'extraction et le calcul de fréquence une première fois sur le corpus complet et ensuite sur chaque sous-corpus par année, en prenant bien soin de distinguer les entités des relations puis de créer des fichiers globaux par concaténation, ce qui nous donne une arborescence de ce type :

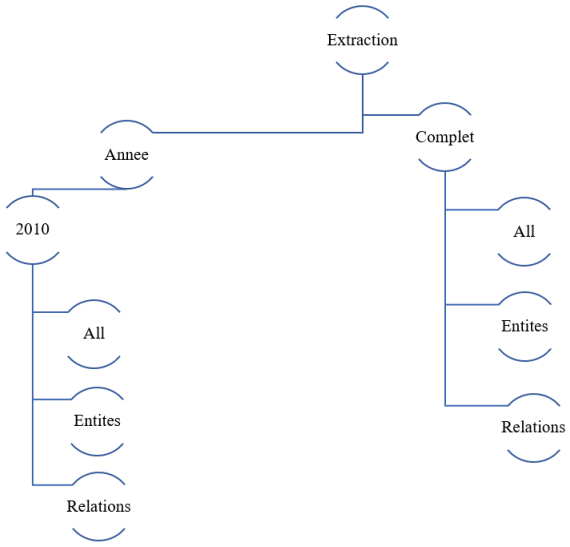

FIGURE 5 – *Arborescence 3*

Modélisation de l'ontologie

Nous avons constaté dans les listes une présence de bruits dus à l'étiquetage (par exemple, *"Donald Trump"* et *"\bDonald Trump"* étaient considérées comme deux entités distinctes). Après avoir nettoyé les fichiers globaux, nous avons supprimé un certain nombre d'entités en ne gardant que celles les plus fréquemment trouvées. Nous nous sommes servi du logiciel de textométrie TXM et de son concordancier pour étudier plus précisément le contexte de ces entités, en commençant par le groupe djihadiste État islamique (EI) :

Contexte gauche	Pivot	Contexte droit
l'offensive spectaculaire des djihadistes de l'Etat islamique (	EI	) dans le pays, c'est au tour de
d'insurgés sunnites menés par les djihadistes ultraradicaux de l'	EI	. Une contre-offensive avait été lancée pour reprendre cet ancien
Irak contrôlée par les djihadistes de l'Etat islamique (	EI	), fuyaient en masse, vendredi 18 juillet,
. « Les dirigeants et les combattants [de l'	EI	] ont beau justifier ces actes abominables par la dévotion
chrétiens menacés par les djihadistes. L'Etat islamique (	EI	) continue à imposer la terreur à une partie de
début de l'offensive des insurgés sunnites menés par l'	EI	qui a précipité le pays dans le chaos et fait
avec les chrétiens et affiché leurs distances vis-à-vis de l'	EI	. Des responsables des villes saintes chiites de Kerbala et
Kurdes réclame des armes pour combattre les djihadistes de l'	EI	Dans une interview à « Bild », dimanche,
qui luttent contre les djihadistes de l'Etat islamique (	EI	). « Nous attendons des armes puissantes non seulement
maintenant commencer à défendre le Kurdistan et à combattre l'	EI	. » M. Barzani a répété au quotidien le
Allemagne. Mais, depuis l'offensive fulgurante de l'	EI	en Irak, au début du mois de juin,
de Mossoul, aux djihadistes de l'Etat islamique (	EI	) qui s'en étaient emparés dix jours plus tôt
détruit ou endommagé dix véhicules armés des insurgés de l'	EI	, sept véhicules de transport Humvee, deux véhicules blindés
l'est du barrage. Un barrage très stratégique L'	EI	, qui s'est emparé de vastes pans du territoire
une lettre au premier ministre irakien Nouri Al-Maliki. L'	EI	utilise les barrages qu'il contrôle comme des armes pouvant
et son barrage est important à l'économie de l'	EI	et à sa volonté de construction d'un Etat incarné
région autonome du Kurdistan après des attaques lancées par l'	EI	début août dans le Nord. A plus de 150
150 km au sud-ouest de Mossoul et alors que l'	EI	accentue sa campagne contre les minorités dans le Nord,
forme nouvelle et très radicale de terrorisme promue par l'	EI	. » Les maisons volées de Kadhafi Sous le régime
du journaliste américain James Foley par l'Etat islamique (	EI	) a fait vivement réagir la communauté internationale. François
de James Foley, Barack Obama a condamné fermement l'	EI	. Le président américain a jugé que le groupe djihadiste
David : Quelle est la relation entre Al-Qaida et l'	EI	? L'EI est l'enfant illégitime d'Al-Qaida,

FIGURE 6 – *Concordancier, TXM*

Ici, nous voyons très clairement apparaitre plusieurs relations, comme par exemple :

— l'EI est l'enfant illégitime d'Al-Qaïda
— (...) Kurdes réclame des armes pour combattre les djihadistes de l'EI
— Barack Obama a condamné fermement l'EI

L'ensemble de ces relations, tissées au fil du corpus, nous a permis de réduire au maximum la liste des entités de manière à ne garder que les principaux acteurs du djihadisme. Après quelques recherches complémentaires, nous avons amorcé un début d'ontologie sur base des principaux groupes djihadistes :

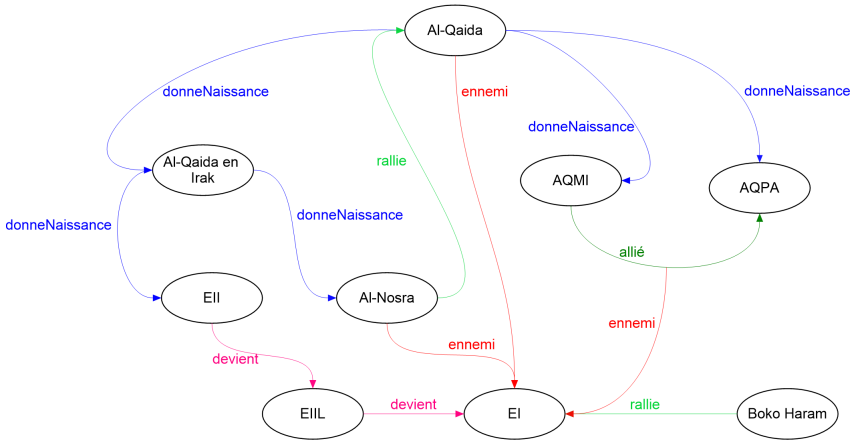

FIGURE 7 – *Ontologie 1*

Analyse critique des résultats

Cette ontologie met en évidence la complexité des relations internationales illustrées. Que penser de l'Arabie Saoudite qui finance l'État Islamique alors qu'elle siège à l'ONU ? Que dire de la présence de la Chine, de la Russie, des États-Unis et de l'Ukraine au Conseil de sécurité de l'ONU lorsqu'on sait les tensions qui agitent ces mêmes pays ? Toutefois, cette ontologie n'a pas pour but d'être exhaustive et certaines situations sont trop complexes pour pouvoir être décrites intégralement dans cette ontologie (notamment la guerre en Syrie). Rappelons-nous aussi que la recherche initiale s'est effectuée à partir du mot "terreur" et non pas "terrorisme" ou "djihadisme", et que nous n'avons travaillé que sur les articles non-payants du journal "Le Monde" et non sur l'intégralité des articles. Enfin, il s'agit ici de corpus issus de la presse et non d'articles scientifiques, ce qui pourrait remettre en cause la pertinence de certaines informations.

Conclusion

Ce projet nous aura permis de construire une ontologie à l'aide d'une méthode bottom-up, de la constitution du corpus à la réalisation de l'ontologie elle-même, en passant par l'étiquetage du corpus et l'extraction des entités nommées. Il serait intéressant d'analyser cette ontologie plus en profondeur et si besoin de la modifier avec l'aide d'un expert dans le domaine du terrorisme et des relations internationales pour ensuite pouvoir la tester sur un autre corpus. D'autres pistes seraient également à explorer, comme par exemple l'analyse comparative des listes d'entités et la construction d'une ontologie pour chaque année afin de mettre en évidence la montée en puissance du terrorisme.

Annexes

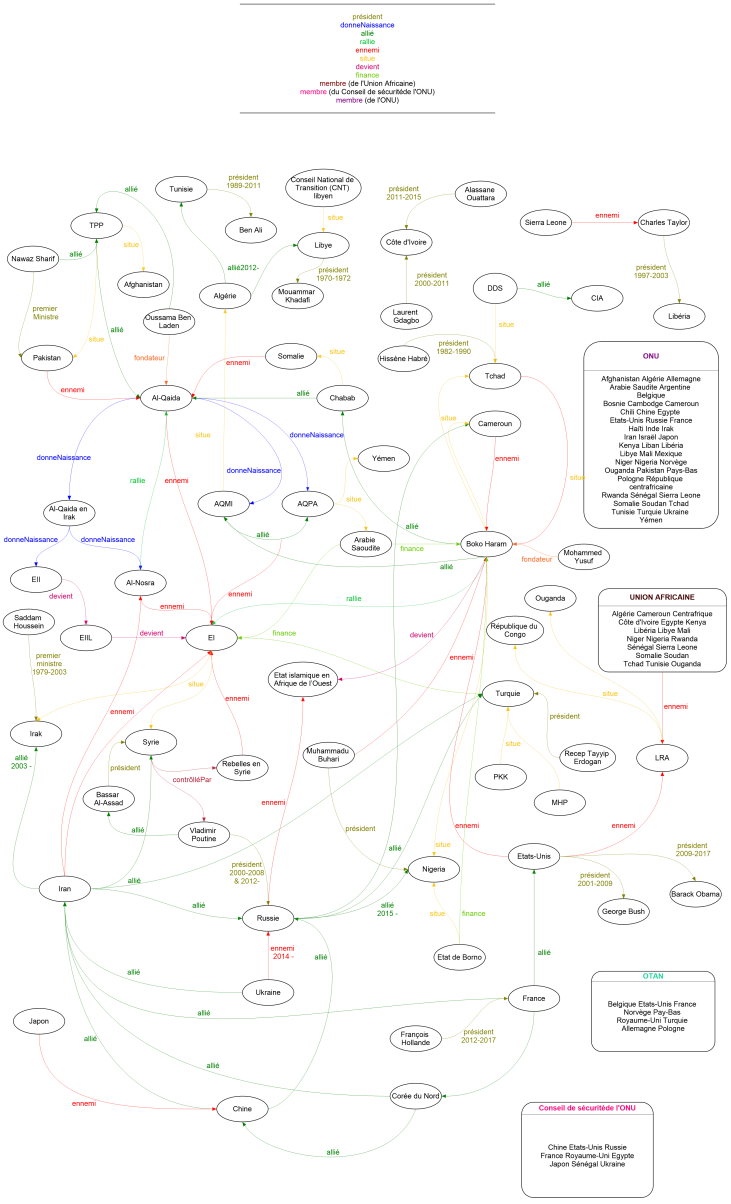

FIGURE 8 – *Ontologie 2 - version simplifiée*

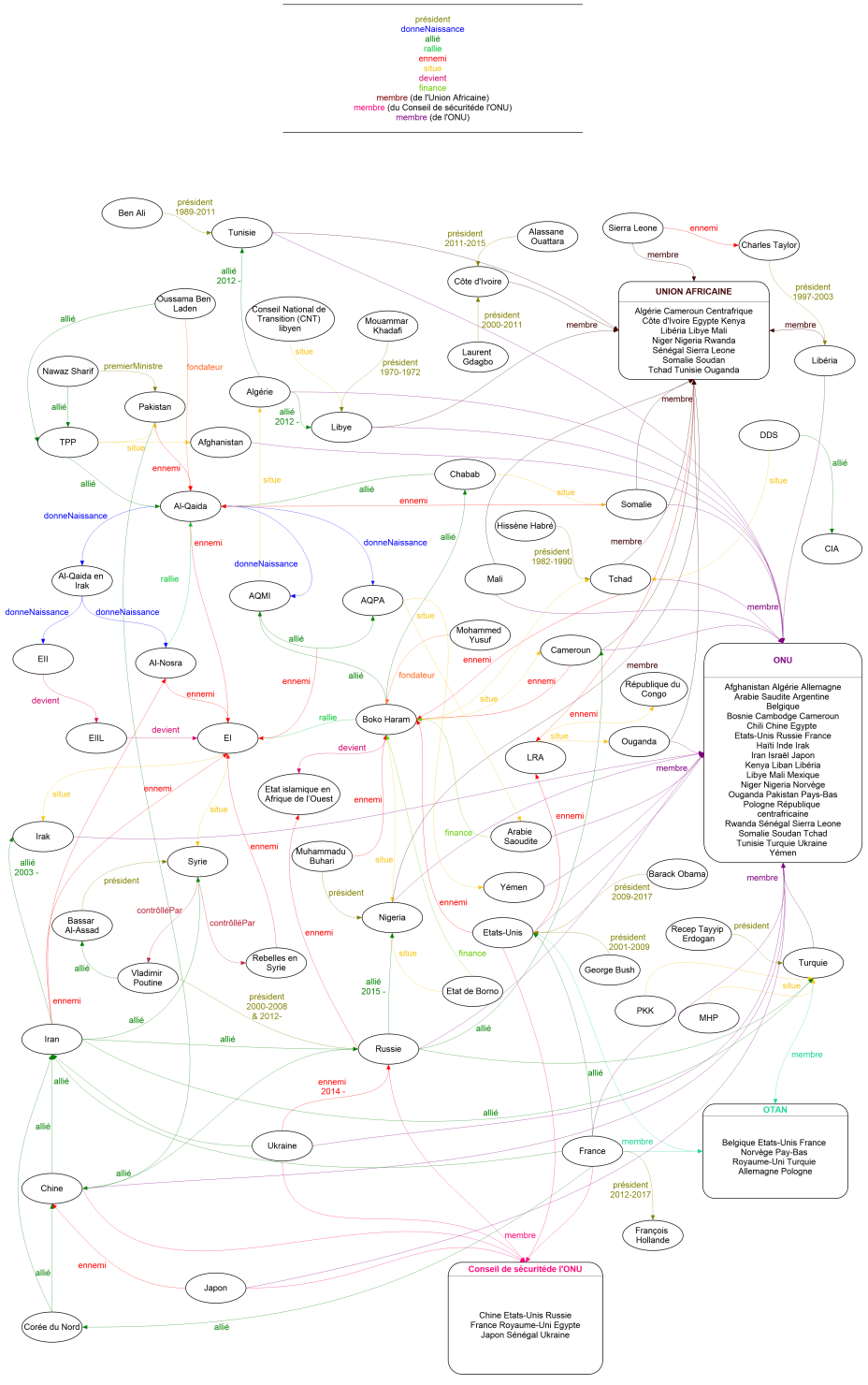

FIGURE *9 – Ontologie 3 - version complexe*

Bibliographie

[1] T. Balzacq. Théories de la sécurité. les approches critiques. *Paris, Presses de Sciences Po*, 2012.

[2] G. Andréani. La guerre contre le terrorisme : un succès incertain et coûteux. *Politique étrangère, 2011/2 (Eté), p. 253-266*, 2011.

[3] C. Guérandel, É. Marlière. Les djihadistes à travers le monde. *Hommes et Migrations 2016/3 (n° 1315). p. 9-16*, 2016.

[4] D. Nouvel, M. Ehrman, S. Rosset. Les entités nommées pour le traitement automatique des langues. *ISTE Editions*, 2015.

[5] A. Rebotier. Développement d'un module d'extraction d'entités nommées pour le français. *Rapport technique, Centre de Recherche Xerox, Grenoble*, 2006.

[6] Br. Tertrais. La "guerre mondiale contre la terreur", 2001-2004. *Politique étrangère, 2004/3 (n° 69), p. 533-546*, 2004.