
بسم الله الرحمن الرحيم
١٤٣٨

Résumé

DBpedia est une ontologie énorme extraite depuis Wikipedia. L'objectif du DBpedia est l'extraction des informations structurées depuis Wikipedia qui est composée principalement des textes. Wikipedia est considérée comme une grande base d'information pour le web et aussi DBepedia est la grande ontologie pour le web sémantique. DBpedia est disponible en plusieurs langues mais malheureusement l'arabe n'est pas totalement disponible.

L'objectif de ce travaille est la conception et l'implémentation d'un algorithme d'extraction des données structurée depuis Wikipedia arabe.

Les pré-requis et moyens indispensables : web sémantique, RDF, Ontologie, Java, Eclipse.

Abstract

DBpedia is a huge ontology extracted from Wikipedia. The goal of DBpedia is the extraction of structured information from Wikipedia which is composed mainly of texts. Wikipedia is considered a great information base for the web and also DBepedia is the great ontology for the semantic web. DBpedia is available in several languages but unfortunately Arabic is not totally available. The goal of this work is the design and implementation of a structured data extraction algorithm from Arabic Wikipedia.

Keywords : web sémantique, RDF, Ontologie, Java, Eclipse.

ملخص

DBpedia هو علم أو نظام ضخم مستخرج من ويكيبيديا, فالهدف من DBpedia هو استخراج المعلومات المنظمة من ويكيبيديا والتي تتكون أساسا من النصوص. تعتبر ويكيبيديا قاعدة معلومات رائعة للويب و أيضا DBepedia هي الأنطولوجيا الرائعة للويب الدلالي. فيتوفر DBpedia بعدة لغات، لكن للأسف اللغة العربية غير متوفرة بشكل كامل.

الهدف من هذا العمل هو تصميم وتنفيذ خوارزمية استخراج البيانات المنظمة من ويكيبيديا العربية.

المتطلبات الأساسية والوسائل الأساسية *web sémantique*:

RDF, Ontologie, Java Eclipse

Remerciement

Nous remercions ALLAH de nous avoir donné la volonté et le courage qui nous a permis de réaliser ce travail, veuille-t-il nous guider dans le droit chemin.

On tient à présenter tous nos respects et notre gratitude à Mr :Djeddai Ala pour son encadrement et pour l'aide qu'il nous a prodigué durant ce mémoire.

On tient, au terme de ce travail, à présenter nos vifs remerciements à tous les personnes qui ont contribué, de près ou loin, à son bon déroulement.

On remercie également les membres des jurys qui ont accepté d'évaluer notre travail.

Table des matières

Liste des figures **8**

Liste des tableaux **10**

1 Définitions générales et présenter les outils d'extraction des données structurées depuis Wikipedia **12**
1.1 Introduction : . . . 13
1.2 Définitions Wikipedia : . . . 14
1.3 L'internationalisation de DBpedia : . . . 15
1.4 DBpedia Ontologie . . . 17
1.5 Modèle d'Infobox . . . 18
1.6 SPARQL . . . 19
1.7 Technologies du Web sémantique : . . . 20
1.7.1 Plateforme de description de ressources (RDF : Ressource Description Framework) . . . 20
1.7.2 Les formats de serialisation . . . 21
1.7.3 Les ontologies : . . . 22
1.7.4 Les langages d'ontologies : . . . 22
1.8 Triplestore : . . . 23
1.9 Un aperçu sur DBpedia : . . . 23
1.10 Lexicalisations : . . . 24
1.11 Wiktionary RDF extraction : . . . 25
1.12 Conclusion : . . . 26

2 Extraction des données structurées depuis wikipédia arabe. **27**
2.1 Introduction : . . . 28
2.2 La Création et Cartographie de modèles arabique en wikipedia : . . . 28
2.3 du Framework DBpediaKnowledge Extraction : . . . 34
2.4 Conclusion : . . . 39

3 Proposer un algorithme d'extraction des données structurées depuis wiki arabe **40**
3.1 Introduction : 41
3.2 Les étapes successives pour accéder au fichier ttl 41
3.3 extraction des données à partir fichier xml : 49
3.4 Créer un algorithme d'extraction : 53
3.4.1 Le diagramme de l'algorithme : 53
3.4.2 Le code avec JAVA : 54
3.5 Conclusion : 59

Bibliographie **61**

Liste des figures

1.1 Une partie des classes d'ontologie DBpedia 17
1.2 Une partie de la description de l'infobox de l'Université 18
1.3 Affiche une requête SPARQL 19
1.4 Extrait de l'ontologie DBpedia. 22

2.1 Mappage du modèle d'infobox 29
2.2 Syntaxe de mappage du modèle d'infobox 30
2.3 Montre qu'il n'y a pas de mapping pour l'arabe Wikipedia plus tôt ce travail 31
2.4 Une partie du modèle d'infobox pour l'infobox de l'université dans l'édition anglaise 32
2.5 Une partie de l'infobox pour l'infobox de l'université en arabe 33
2.6 infoboxHouari_Boumédiènne 36
2.7 étiquettes pour la page de Houari_Boumédiènneviardfs : label 36
2.8 exmple code infbox de mois 37
2.9 exemple code infbox de pays 38
2.10 exemple code infobox de page web 39

3.1 Placez les deux dossier dans d :/ 41
3.2 créationdes variables d'environnement 42
3.3 ajouter les valeurs dans la variable (path) 43
3.4 site d'exportation wikipédia arabe 45
3.5 enregestré page xml à partir le site 47
3.6 Lorsque nous ouvrons le fichier par nodepad++ , nous trouvons ce qui suit 47
3.7 Fichier xml aprés l'exportation 48
3.8 Extraction default properties 49
3.9 extraction les fichiers ttl et tql 50
3.10 Contenu de fichier ttl 51
3.11 contenu de fichier ttl (moète Mohammed Chebaiki) 52
3.12 Le diagramme de l'algorithme 53

3.13 Code java de téléchcharger fichier xml 54
3.14 télécharger le fichier xml 55
3.15 Enregestré le fichier xml dans le chemin :... \20180301 56
3.16 code java pour d'extraction le fichier xml 57
3.17 l'emplacement de fichier ttl 58

Liste des tableaux

1.1 Exemple :(Sujet,Prédicat,Objet) 21

Introduction générale :

L'arabe est considéré comme l'une des langues les plus anciennes de l'humanité et son plus profond en termes de significations et de concepts spéciaux est la langue de la langue coranique du miracle linguistique Avec l'avantage de la langue arabe nous avons suggéré ce sujet Extraction des données Structurées depuis Wikipédia Arabe

qui étudiera comment importer des données du fichier XMLt pour obtenir fichier RDF en proposant un algorithme spécial pour cela, nous avons divisé notre étude en 3 chapitres du mémorandum :

Chapitre 1 :Vue d'ensemble des définitions du sujet en plus des définitions des programmes utilisés et liés à Wikipédia, arabe ou non-arabe, qui sont inclus dans l'environnement de Wikipedia, en plus de quelques exemples de ces programmes.

Chapitre 2 : Nous fournissons un exemple de l'importation de données à partir de Wikipédia arabepour expliquer les composants de Wikipedia et donner une explication et fournir une ventilation sur certains composants.

Chapitre 3 :Nous avons présenté en détail l'algorithme proposé et comment importer les données en arabe et fournir les étapes d'importation de données que nous avons proposées pour obtenir le fichier de données rdf.

Chapitre 1

Définitions générales et présenter les outils d'extraction des données structurées depuis Wikipedia

1.1 Introduction :

Wikipedia est l'une des plus grandes sources de connaissance de l'humanité et la plus grande encyclopédie sur le Web en étant le 7ème site le plus visité selon alexa.com (en juillet 2011).

Wikipedia est créé uniquement par la création collaborative de ses utilisateurs. Wikipedia est disponible dans plus de 280 langues, et la Wikipédia en anglais contient plus de 3,5 millions [1] articles (en juillet 2011) a. Cependant, malgré son succès, il y a aussi quelques problèmes et inexploitéspotentiel :

— les capacités de recherche de Wikipedia sont limitées à la correspondance des mots clés.

— des incohérences peuvent survenir en raison de la duplication d'informations sur des pages différentes.et dans différentes éditions de la Wikipédia.

1. http ://fr.wikipedia.or/wiki/wikipedia :Size_of_wikipedia

1.2 Définitions Wikipedia :

Est un projet d'encyclopédie universelle, multilingue, créé par Jimmy Wales et Larry Sanger le 15 janvier 2001 en wiki sous le nom de domaine wikipedia.org. Les versions des différentes langues utilisent le même logiciel de publication, MediaWiki, et ont la même apparence, mais elles comportent des variations dans leurs contenus, leurs structures et leurs modalités d'édition et de gestion.

L'encyclopédie ainsi que son environnement sont hébergés sur Internet grâce aux serveurs financés par la WikimediaFoundation, organisation à but non lucratif américaine, dépositaire de la marque Wikipédia.

L'encyclopédie est en libre accès, en lecture comme en écriture, c'est-à-dire que n'importe qui peut, en accédant au site, modifier la quasi-totalité des articles publiés sous licence CC-BY-SA 3.0. En octobre 2017, Wikipédia est le cinquième site le plus fréquenté au monde1, et constitue le plus grand et le plus populaire des ouvrages de références générales d'Internet. En février 2014, il y avait mondialement presque 500 millions de visiteurs chaque mois et en mai 2014, 22 millions de comptes pour les éditeurs, dont plus de 73 000 sont considérés comme actifs. [1]

Historiquement, l'anglais a été la principale langue utilisée, avant qu'une multitude de sites ne soient ouverts dans d'autres langues (283 langues en 20162). La Wikipédia en anglais a toujours conservé cette importance relative : le nombre d'articles s'élève à plus de 5 millions.

Tim Berners Lee a défini le Web sémantique, car il permet non seulement de travailler en coopération avec les machines et les personnes, mais aussi de structurer et de donner du sens au contenu Web, ce qui facilite l'automatisation des tâches et les requêtes complexes. Mais, malheureusement, l'image totale du Web sémantique n'est pas terminée jusqu'à présent. Certains efforts et projets tentent d'utiliser les technologies du Web sémantique pour enrichir structurellement et sémantiquement le Web. L'un de ces projets est le projet DBpedia, fruit de la collaboration entre l'Université Freie de Berlin, l'Université de Leipzig et le logiciel Open Link. Le projet DBpedia, qui a débuté en 2007, vise à extraire des informations structurées de Wikipédia en tant que triplets RDF pour faciliter l'exécution de requêtes sophistiquées pour les applications sémantiques.

1. http ://fr.wikipedia.or/wiki/wikipedia :Size_of_wikipedia

1.3 L'internationalisation de DBpedia :

Actuellement, au moment de la rédaction de cet article, l'ensemble de données DBpedia décrit plus de 3,64 millions de choses dans 97 langues fournissant 1 milliard d'informations sous la forme de triplets RDF avec une licence ouverte couvrant différents domaines, tels que ; personnes, organisations, logiciels, films, livres, etc..[1]

Un nombre croissant d'éditeurs, contributeurs et fournisseurs de données publient leurs données sur DBpedia en supposant que DBpedia améliore et améliore la qualité des données et recherche dans le contenu textuel de Wikipédia selon la capacité de DBpedia à extraire des informations structurées de Wikipédia et à les mettre à jour régulièrement si des changements se produisent dans Wikipedia.

Une version antérieure de DBpedia était en anglais. DBpedia Information Extraction Framework (DIEF) aide DBpedia à élargir sa portée et à couvrir d'autres langues de Wikipedia en ajoutant de nouveaux chapitres.

La version disponible de DBpedia au moment de la rédaction de ce document couvre 15 chapitres localiser DBpedia qui comprennent les langues suivantes : catalan, allemand, grec, espagnol, français, irlandais, croate, hongrois, italien, néerlandais, polonais, portugais, russe, slovène, turc en plus de l'anglais. [1]

Malheureusement, la langue arabe n'était pas disponible dans ces chapitres localisés de DBpedia. Cependant, le nombre d'internautes parlant l'arabe est de 86 077 806, soit environ 3,8% des internautes dans le monde, ce nombre est relativement important si on le compare au nombre d'utilisateurs qui parlent d'autres langues.

D'autre part, ce grand nombre d'utilisateurs n'a pas été reflété réellement sur le nombre total d'articles disponibles dans l'édition arabe, qui est d'environ 179.044 article du contenu de Wikipedia, si on le compare au nombre total d'articles, qui est d'environ 22.175.193 disponibles en 285 langues. L'édition arabe a été classée comme la 25th. [2]

1. http ://www.gov.de/wikis/cls/rnd_in_rdf
2. http ://navigator.dbpedia.org

la plus grande édition de Wikipedia par un pourcentage de 0.81% du nombre total d'articles dans Wikipedia. Cependant, le nombre d'utilisateurs enregistrés sur Wikipédia en arabe est estimé à 0,59% du nombre total d'internautes arabophones.

Chercher les raisons derrière la scène pour le faible nombre d'articles publiés en langue arabe dans Wikipedia est au-delà de la portée de cette recherche, Cependant, il est prévu que le nombre d'articles dans le domaine arabe sur Wikipedia peut augmenter après le déploiement de l'arabe DBpedia selon la capacité de DBpedia à faire des choses qui n'auraient pas pu être faites auparavant dans Wikipedia.

Grâce à cette section, nous décrirons brièvement l'ontologie DBpedia, l'idée au-delà du modèle infobox, le langage SPARQL considéré comme un protocole et un langage de requête utilisé pour extraire des données de DBpedia via des points de terminaison tels que Virtuso et Snorql.

1.4 DBpedia Ontologie

L'Ontologie DBpedia est créée manuellement en fonction du chapitre anglais des infoboxes Wikipédia. Le nombre de classes qui forment l'ontologie est passé de 170 classes à 359 classes en plus de la croissance du nombre de propriétés représentant différentes relations avec le domaine et va de 720 à 1775 eux :[2]



FIGURE 1.1 – Une partie des classes d'ontologie DBpedia [9]

1.5 Modèle d'Infobox

Il est considéré comme une forme tabulaire placée dans le coin supérieur de l'article de Wikipedia décrivant les attributs de cet article qui résument l'article sous la forme d'attributs-valeur.

L'infobox est entourée d'opérateurs {{}}. La figure 2 montre le code généré et la description de l'infobox de l'université qui contient le nom de l'université, le type, le président et d'autres attributs nécessaires pour expliquer tout modèle de type université.[3]

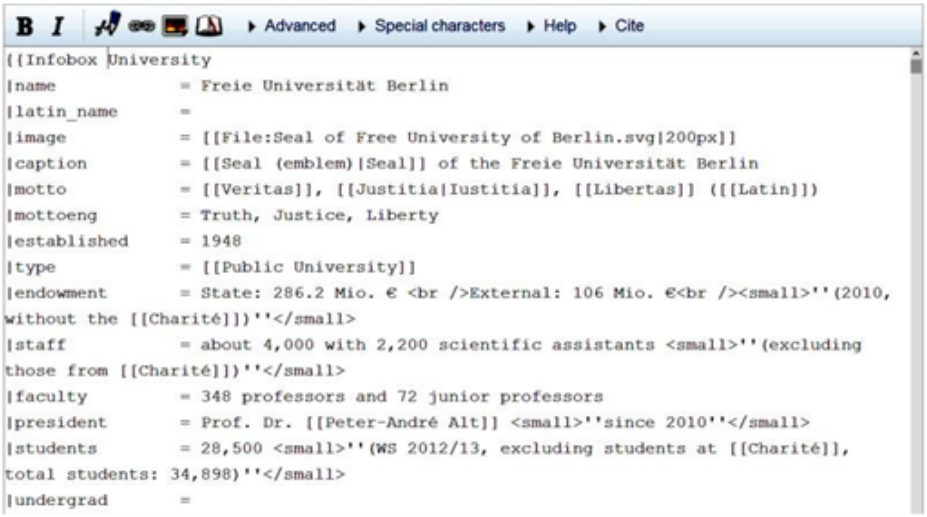


FIGURE 1.2 – Une partie de la description de l'infobox de l'Université

1.6 SPARQL

SPARQL est considéré comme le langage de requête pour le Web sémantique et est utilisé comme point final pour interroger la base de connaissances DBpedia afin de récupérer des triplets qui peuvent aider au développement d'applications basées sur les informations structurées de DBpedia. Pour interroger des informations de DBpedia n'est pas un processus sophistiqué, c'est comme les autres langages de requête, la seule différence dans la forme de l'instruction qui dépend de l'idée de représenter des données sur des triplets. SPARQL est devenu une norme en 2008. [4] et à partir de ce moment, il est utilisé dans la plupart des applications Web sémantique en raison de sa flexibilité pour récupérer des informations à partir du Web de données. [1]



FIGURE 1.3 – Affiche une requête SPARQL

1. http ://dbpedia.org/sparql

1.7 Technologies du Web sémantique :

Tim Berners-Lee définit le Web sémantique en écrivant : « Le Web sémantique n'est pas un Web séparé, mais une extension de l'existant, dans laquelle on donne à l'information un sens bien défini permettant aux ordinateurs et aux gens de mieux travailler en coopération». [1]

Selon le consortium du World Wide Web (W3C), le concept peut être défini comme suit : « Le Web sémantique est l'idée d'avoir les données sur le web définies et liées de façon qui peut être utilisée par les machines non seulement pour des fins d'affichage, mais également pour l'automatisation, l'intégration et la réutilisation des données à travers des applications diverses».

Pour réaliser cette vision, différentes technologies ont été développées comme : RDF,RDFS,OWL,.....etc

1.7.1 Plateforme de description de ressources (RDF : Ressource Description Framework)

RDF est un langage de description de connaissances utilisé pour décrire des ressources variées comme les ressources Web, les entités et les concepts. RDF est pour le Web sémantique ce qu'est HTML pour le web conventionnel.

RDF utilise différents éléments : les ressources (appelées également entités), les propriétés (appelées également prédicats ou rôles) et les expressions (appelées triplets ou faits).

Les ressources incluent n'importe quelle entité, concrète ou abstraite. Chaque ressource est identifiée par un identificateur uniforme de ressource (URI : Uniform Ressource Identifier) qui doit être unique.

Exemple : [2]

Les propriétés sont des ressources spécifiques qui peuvent être utilisées pour décrire les attributs ou les relations d'autres ressources.
Exemple : [3]

1. http ://wiki.dbpedia.org/developer_Documentation/Extractor
2. *http : //dbpedia.org/resource/Williamshakespeare*
3. *http : //dbpedia.org/ontology/Person/height*

Les expressions sont utilisées pour décrire les propriétés des ressources sous le format : sujet, prédicat, objet. Les trois composants peuvent être des ressources (URI). Le sujet et l'objet peuvent êtres des nœuds vides.

TABLE 1.1 – Exemple :(Sujet,Prédicat,Objet)

Subject	**Predicate**	**Object**
dbpedia :William_Shakespeare	dbp :birthplace	"Stratford-upon-Avon, Warwickshire England"
dbpedia :William_Shakespeare	dbp :dateOfBirth	"April 1564"
dbpedia :William_Shakespeare	dbp :occupation	"Playwright, poet, actor"
dbpedia :William_Shakespeare	dbpo :child	dbpedia :Judith_Quiney
dbpedia :Judith_Quiney	dbp :spouse	"Thomas Quiney"

1.7.2 Les formats de serialisation

Chaque triplet peut être écrit sur une ligne se terminant par un point. Le fichier a généralement l'extension ".nt". Il y a d'autres formats comme celui utilisant la notation XML (RDF/XML), N3 et Turtle.[10]

1.7.3 Les ontologies :

Selon le W3C, « Une ontologie définie les termes utilisés pour décrire et représenter un domaine de connaissance. »

Exemple [5] :

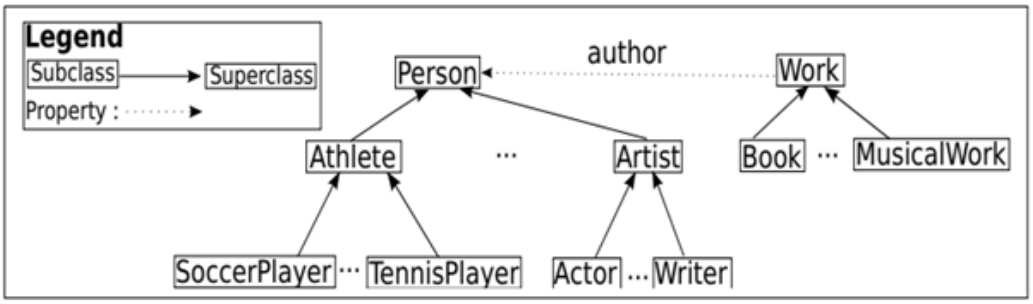


FIGURE 1.4 – Extrait de l'ontologie DBpedia.

1.7.4 Les langages d'ontologies :

Les ontologies peuvent être crées en utilisant différents langages parmi lesquels les schémas RDF (RDFS) et l'OWL (Ontology Web Language).

RDFS est un langage relativement simple qui peut être utilisé pour créer un vocabulaire pour décrire les classes, les sous-classes et les propriétés des ressources RDF. OWL est basé sur RDFS et permet d'exprimer des relations plus complexes.

Exemple : RDFS

1. dbo :Writerrdf :typeowl :Class .
2. dbo :Writerrdfs :subClassOfdbo :Artist .
3. dbo :authorrdf :typeowl :ObjectProperty .

Exemple : Qui est l'épouse du fils de William Shakespeare

1. PREFIX dbp : <http ://dbpedia.org/property/>
2. PREFIX dbpedia : <http ://dbpedia.org/resource/>
3. PREFIX dbo : <http ://dbpedia.org/ontology/>
4. SELECT ?spouse WHERE {

5. $dbpedia : William_Shakespearedbo : child?child.$
6. $?childdbp : spouse?spouse.$
7. }.

1.8 Triplestore :

C'est un programme capable de stocker est indexer les données RDF pour pouvoir les interroger efficacement. La plupart des triplestores supportent le langage SPARQL.

1.9 Un aperçu sur DBpedia :

Le cœur de DBperdia est un processus d'extraction de boites d'informations (infobox) qui sont des modèles que contiennent plusieurs articles Wikipédia. Les infoboxes affichent les faits les plus pertinentes d'un article Wikipédia dans un tableau souvent sous forme de paires attribut-valeur.

1.10 Lexicalisations :

La lexicalisation est définie par WordNet comme «le processus de faire un mot pour exprimer un concept» . Dans le contexte de ce projet, les lexicalisations sont des formes de surface faisant référence à une ressource DBpedia donnée. Le jeu de données DBpediaLexicalizations stocke les relations entre les ressources DBpedia et un ensemble de formulaires de surface que nous avons trouvé référents à ces ressources dans Wikipedia. [1]

Nous utilisons le graphe des étiquettes, des redirections et des désambiguations dans DBpedia pour extraire un lexique qui associe plusieurs formes de surface à une ressource et interconnecte plusieurs ressources à une étiquette ambiguë.

Les étiquettes des ressources DBpedia sont créées à partir des titres de pages Wikipedia, qui peuvent être considérés comme des formulaires de surface approuvés par la communauté. Les redirections vers les URI indiquent des synonymes ou des formes de surface alternatives, y compris des fautes d'orthographe et des acronymes communs. Leurs étiquettes deviennent également des formes de surface.

Les désambiguïsations fournissent des formes de surface ambiguës qui sont «confuses» avec toutes les ressources auxquelles elles sont liées. Leurs étiquettes deviennent des formes de surface pour toutes les ressources cibles de la page de désambiguïsation. Notez que nous effaçons le suivi

Entre parenthèses à partir des étiquettes lors de la construction des formes de surface. Par exemple l'étiquette / Copyright (bande) / produit la forme de surface "Copyright". Cela signifie que les étiquettes de ressources et de redirections peuvent également introduire des formes de surface ambiguës, en plus des étiquettes provenant des titres des pages de désambiguïsation.

La collection de formes de surface créée en conséquence constitue un ensemble contrôlé d'étiquettes communément utilisées pour les ressources cibles. Une autre source de références textuelles aux ressources DBpedia est wikilinks. Nous utilisons les wikilinks pour estimer la probabilité que des formes de surface se référant à un ressources candidates Nous considérons chaque wikilink comme une preuve que le texte d'ancrage est un formulaire de surface couramment utilisé pour la ressource DBpedia représentée par la

1. http ://creativecommons.org

cible du lien. En comptant le nombre de fois qu'une forme de surface s'est produite avec une ressource DBpedia, nous pouvons estimer un score qui indique le niveau d'association entre la forme de surface et la ressource.

Le jeu de données DBpediaLexicalizations est produit par l'équipe DBpediaSpotlight. Il est utilisé pour créer l'index derrière le service DBpediaLookup, ainsi que les mappages candidats derrière DBpedia Spotlight..[1]

1.11 Wiktionary RDF extraction :

Wiktionnaire, le dictionnaire libre, est un autre projet de la Wikimedia-Foundation d'où DBpedia extrait des données RDF structurées. Tout comme Wikipédia, Wiktionary est disponible dans de nombreuses langues, telles que Wiktionary en anglais (http ://en.wiktionary.org) et Wiktionary en allemand (http ://de.wiktionary.org). Cependant, chacun de ces sites indépendants contient des entrées dans plusieurs langues.

Pour le mot français deux, il y a une entrée dans le Wiktionnaire anglais (http ://en.wiktionary.org/wiki/deux) et une autre entrée dans le Wiktionnaire allemand (http ://de.wiktionary.org/wiki/deux).

Pour un mot comme dans, une seule page wiki dans le Wiktionnaire en anglais (http ://en.wiktionary.org/wiki/in) contient des sections pour 24 langues différentes qui utilisent ce mot. (La structure exacte des entrées diffère entre les langues de Wiktionary, et légèrement entre les langues de chaque site, de sorte que les entrées danoises et néerlandaises dans le Wiktionnaire anglais peuvent utiliser différents types de modèles de wiki. [2]

Nous visons à fournir un framework open-source (basé sur DBpedia) pour extraire des ressources lexicales sémantiques (une ontologie sur l'utilisation du langage) de Wiktionary.

Les données incluent actuellement la langue, la partie du discours, les sens, les définitions, les synonymes, les taxonomies (hyponymes, hyperonymes, synonymes, antonymes) et les traductions pour chaque mot lexical. L'accent est mis sur la flexibilité (pour le schéma libre) et la configurabilité (vers différentes versions de Wiktionary).

1. http ://creativecommons.org
2. http ://www.gdv.de/wikis/cls/PND_in_RDF

La configuration est effectuée dans un fichier XML qui code les mappages de langue et les modèles d'analyse contenant des espaces réservés. L'objectif est de permettre l'ajout de langues simplement par configuration sans avoir besoin de compétences en programmation et sans altérer le code source de Scala. Cela permet une publicité rapide et bon marché pour de nouveaux scénarios d'utilisation.

Les données extraites peuvent (en raison de leur richesse sémantique) être automatiquement transformées en modèle Lemon ou en formats spécifiques au domaine plus simples - par exemple une représentation CSV de la traduction, qui peut être chargée dans une base de données relationnelle.

En offrant un service de données liées, nous espérons étendre le rôle central de DBpedia dans l'infrastructure LOD au monde de la linguistique ouverte. La décharge RDF contient actuellement 100 millions de triplets.

1.12 Conclusion :

Dans le chapitre, nous présentons une définition des différents termes utilisés dans la base de données Wikipédia et un aperçu général du sujet, ainsi qu'une image générale de Wikipédia contenant les composants, programmes et applications qui aident Wikipédia et Wikipedia à illustrer ces éléments sous forme d'images.

Chapitre 2

Extraction des données structurées depuis wikipédia arabe.

2.1 Introduction :

Dans ce chapitre, nous parlerons des caractéristiques de Wikipédia en arabe et de ses avantages en fournissant des exemples et en expliquant certaines caractéristiques La langue arabe est incluse dans Wikipedia Comme nous savons que l'arabe est une langue riche en significations et concepts, nous trouvons à peine un terme arabe mais il a au moins deux ou plusieurs significations.

2.2 La Création et Cartographie de modèles arabique en wikipedia :

DBpedia arabe, aime les autres chapitres de DBpedia, a quelques problèmes dans les mappages ; l'un de ces problèmes est l'indisponibilité des infoboxes pour de nombreux articles arabes. Par exemple, dans la liste des musées égyptiens qui considèrent les pages de valeur dans le patrimoine culturel, il n'y a que 4 musées sur 50 qui ont des infoboxes, ce qui est un pourcentage relativement faible, 8% de la liste complète.[6]

Pour cette raison ; de nombreuses infoboxes pour divers articles ont été construites, mais cela nécessite plus d'efforts de la part des contributeurs et de la communauté pour en finir autant que possible pour augmenter la quantité de données extraites du Wikipedia arabe alimentant le chapitre arabe de Dbpedia.

En raison de l'importance du modèle d'infobox pour DBpedia, et parce que le travail ici est considéré comme la première tentative dans le déploiement de la DBpedia arabe, cet article expliquera simplement les différentes étapes nécessaires pour la cartographie à partir de zéro. Le processus commence avec la création de l'infobox comme un effort pour augmenter le nombre d'infoboxes dans les articles écrits en langue arabe.

Pour vérifier le mappage d'un modèle d'infobox, tout ce dont nous avons besoin est de prendre le nom d'infobox concaténé avec l'URI du chapitre arabe dans les statistiques de mappage, par exemple: mappings.dbpedia.org/index.php/mapping_ar: "معلوماتجامعة "

.



Mapping ar:معلومات جامعة

This is the mapping for the Wikipedia template معلومات جامعة. Find usages of this Wikipedia template here.

Test this mapping with some example Wikipedia pages. Check which properties are not mapped yet.

Read more about mapping Wikipedia templates.

Template Mapping (help)	
map to class	University

Mappings

Property Mapping (help)	
template property	الإسم
ontology property	foaf:name

Property Mapping (help)	
template property	الشعار
ontology property	motto

Property Mapping (help)	
template property	التأسيس
ontology property	foundingDate

Property Mapping (help)	
template property	الرئيس
ontology property	chairman

FIGURE 2.1 – Mappage du modèle d'infobox

Si l'infoboxWikipedia n'a pas été mappée auparavant, la page de mappage dans DBpedia peut être créée en mappant le nom de l'infobox et les attributs aux propriétés de classe appropriées à l'intérieur de l'ontologie, comme illustré à la Figure suivant .

```
{{TemplateMapping
| mapToClass = University
| mappings =
{{PropertyMapping | templateProperty = الإسم | ontologyProperty
foaf:name }}
{{PropertyMapping | templateProperty = الشعار | ntologyProperty
motto }}
{{PropertyMapping | templateProperty = التأسيس | ntologyProperty
= foundingDate
}}
{{PropertyMapping | templateProperty = الرئيس | ontologyProperty
= chairman }}
{{PropertyMapping | templateProperty = العميد | ontologyProperty
= chancellor }}
{{PropertyMapping | templateProperty = عميد الكلية | ontologyProperty
= president }}
{{PropertyMapping | templateProperty = النائب | ontologyProperty =
vicePresident }}
```

FIGURE 2.2 – Syntaxe de mappage du modèle d'infobox

Chaque infobox est mappée à une classe et chaque attribut est mappé à une propriété dans l'ontologie Dbpedia.Infobox est considéré comme un drapeau indiquant que cette page Wikipedia appartient à une classe particulière dans l'ontologie DBpedia et ce regard dans le attribut mapToClass ; tandis que l'attribut seconds mappings représente le mappage des attributs aux propriétés à l'intérieur de l'ontologie DBpedia.

La langue arabe est considérée comme une langue non-latine qui contient des caractères qui apparaissent comme des symboles via les navigateurs web selon ASCII, c'est pourquoi les utilisateurs sont invités à utiliser Mozilla Firefox ou Google Chrome lorsqu'ils traitent avec la DBpedia arabe.

Avant le travail effectué dans ce document, le chapitre arabe de DBpedia n'a pas encore été discuté et il n'y avait pas de chapitre arabe dans DBpedia. Après la création d'un arabe namespace "ar" sur le wiki, à la demande de l'auteur de cet article, les occurrences de propriété et de modèleétaientnulles au début ; iln'y a pas de cartographiecommedans la figure suivant . Après le travail pratique de cetterecherche ; Le nombre de classes et de propriétéscartographiéesaugmentemaintenant pour atteindre 17,63%, commedans la figure suivant , et devrait augmenter si de nouveaux contributeurs et volontaires se joignent au travail de la DBpedia arabe.1

occurrences template (with link to property statics)			num properties mapped properties (%)		num property occurrences mapped property occurrences (%)	
4273	تصنيف كائن	Edit	50	0.00	63796	0.00
2245	infobox setlement	Edit	330	0.00	47910	0.00
2042	معلومات شخصية	Edit	49	0.00	14727	0.00
2031	معلومات بلدة ايطاليه	Edit	105	0.00	44034	0.00
1527	معلومات فيلم	Edit	33	0.00	23682	0.00
1509	معلومات شخصية كرة قدم	Edit	25	0.00	23254	0.00
1298	قرية مصرية	Edit	17	0.00	7270	0.00
1292	معلومات ممثل عربى	Edit	27	0.00	7444	0.00
1227	معلومات مقاطعة امريكية	Edit	43	0.00	6784	0.00
1156	معلومات ممثل	Edit	21	0.00	6925	0.00
855	معلومات شركة	Edit	28	0.00	8633	0.00
855	معلومات فنان	Edit	24	0.00	9689	0.00
705	معلومات كيمياء جديدة	Edit	59	0.00	2255	0.00
701	تصنيف علمى	Edit	26	0.00	5120	0.00
671	معلومات مسلسل	Edit	26	0.00	9193	0.00

FIGURE 2.3 – Montre qu'il n'y a pas de mapping pour l'arabe Wikipedia plus tôt ce travail

Pour chaque article dans Wikipedia il ya un parapluie général couvre chaque sujet, par exemple, le parapluie qui couvre l'infobox de l'Université Larbi Tébessi " جامعة العربي تبسي" est "Université". Pour afficher le modèle général pour University, ouvrez une nouvelle fenêtre pour la version anglaise de cette infobox via: http://en.wikipedia.org/wiki/Template:Infobox_University comme dans la Figure suivant .

```
{{Infobox university
| name                =
| native_name         =
| image_name          =
| image_size          =
| image_alt           =
| caption             =
| latin_name          =
| motto               =
| motto_lang          =
| mottoeng            =
| established         = <!-- {{Start date|YYYY}} -->
| closed              = <!-- {{End date|YYYY}} -->
| type                =
| affiliation         =
| endowment           =
| budget              =
| officer_in_charge   =
| chairman            =
| chancellor          =
```

FIGURE 2.4 – Une partie du modèle d'infobox pour l'infobox de l'université dans l'édition anglaise

Dans le menu de gauche, le modèle arabe pour cette infobox peut être sélectionné, et les attributs de " جامعةالعربي تبسي" peuvent être remplis par ses valeurs, tandis que d'autres attributs inconnus peuvent être laissés vides jusqu'à ce que d'autres contributeurs les remplissent plus tard.



FIGURE 2.5 – Une partie de l'infobox pour l'infobox de l'université en arabe

Il existe une autre façon de voir le modèle d'infobox pour "جامعة" directement en écrivant:

"ar.wikipedia.org/wiki/Template: جامعة"Ou "wikipedia.org/wiki/ قالب: جامعة"

Comme indiqué précédemment, il y a un certain nombre de problèmes survenus dans le chapitre arabe de DBpedia, y compris l'indisponibilité des infoboxes discutés précédemment; en plus des problèmes survenus lors de la cartographie et l'extraction des entités.

des problèmes dans Wikipédia en arabe, est l'apparition de noms différents pour le même attribut, comme l'attribut date de naissance qui apparaît dans différentes infoboxes sous différents noms: une fois comme "تاريخالميلاد" une autre fois comme "ولدفي ", la troisième fois comme" تاريخالولادة "et la quatrième fois comme" ولد ". Mais ces conflits pour tous ces attributs peuvent être facilement mis en correspondance avec la date de naissance dans l'ontologie

DBpedia http ://dbpedia.org/ontology/birthDate et De cette façon, nous pouvons sauter sur différents noms pour le même attribut

Un autre problème survenu en arabe DBpedia peut être conclu comme l'apparition de divers noms d'infobox faisant référence au même concept tel que :

Ce problème peut être résolu en mappant les quatre concepts à la même classe d'ontologie équivalente, qui est la classe Musical Artist dans l'ontologie DBpedia.

Incohérence des noms entre l'infobox et son modèle considéré aussi l'un des problèmes dans le chapitre arabe de DBpedia, par exemple, il y a un modèle appelé "جامعة" tandis que le nom d'infobox est appelé "معلومات جامعة" et pour la facilité d'utilisation, nous recommandons que le nom de l'infobox et le nom du modèle aient le même nom.

Dans la plupart des modèles arabes, lorsque les noms géographiques sont placés dans l'infobox, ils créent des problèmes dans le formatage des modèles, mais lorsqu'ils sont situés en dehors de l'infobox, ils fonctionnent sans problème dans Wikipedia, mais dans ce cas ils deviennent inutilisables à DBpedia.

2.3 du Framework DBpediaKnowledge Extraction :

L'architecture du cadre d'extraction sera décrite dans cette section montrant les tâches de divers extracteurs et les modifications qui peuvent faciliter l'extraction d'articles écrits en langue arabe.

— **Label Extractor :** Cet extracteur est utilisé pour extraire le titre de l'article dans diverses éditions de Wikipedia et les placer dans rdfs : label. Un exemple pour l'extracteur d'étiquette est la page du célèbre Le défunt président Houari Boumediene dans Wikipedia qui a cette URL :

"https ://ar.wikipedia.org/wiki/%D9%87%D9%88%D8%A7%D8 %B1%D9%8A_%D8%A8%D9%88%D9%85%D8%AF%D9%8A %D9%86 ?action=edit&veswitched=1&oldid=28331318

" Cet article est disponible en 14 langues différentes, cependant, la page de Houari_Boum dans DBpedia via "dbpedia.page/Houari_Boumédiènne" n'a que 5 valeurs pour le rdfs : label pour seulement 5 langues, pas pour toutes les 14 langues utilisées pour la publication de cet article. Les 5 seules langues extraites sont les langues qui ont des chapitres localisés dans DBpedia dans leur langue locale et ont des extracteurs différents, dont l'extracteur titre d'un article dans divers chapitres de Wikipédia et placez-le sur rdfs : label dans l'ontologie DBpedia. De Houari_Boumédiènne.



FIGURE 2.6 – infoboxHouari_Boumédiènne

```
{{معلومات صاحب منصب
| اسم          = هواري بومدين
| الصورة       = Houari Boumediene's Portrait.jpg
| المنصب       = {{رمز علم|الجزائر}} [[قائمة رؤساء الجزائر|الرئيس الثاني]] [[الجزائر|الجمهورية
الجزائرية]]
| بداية الفترة = [[19 يونيو]] [[1965]]
| نهاية الفترة = [[27 ديسمبر]] [[1978]]
| سبقه         = [[أحمد بن بلة]]
| تبعه         = [[رابح بيطاط]]
| المنصب1      = [[وزارة الدفاع الوطني (الجزائر)|أول وزير الدفاع]]
| الرئيس1      = *[[أحمد بن بلة]]
* هواري بومدين
| النائب1      =
```

FIGURE 2.7 – étiquettes pour la page de Houari_Boumédiènneviardfs : label

Les exemples des Infobox du page Wikipédia de Houari _Boumédiènne

- **Infobox de mois :**



FIGURE 2.8 – exmple code infbox de mois

• Ce modèle contient des informations utilisateur sur le nom, la définition, la date et l'heure de la modification, ainsi que le type de texte et le format du texte.

- **l'infobox de pays :**


```
<page>
  <title>قالب:بيانات بلد الجزائر</title>
  <ns>10</ns>
  <id>295808</id>
  <revision>
    <id>22334127</id>
    <parentid>22332032</parentid>
    <timestamp>2017-01-31T03:38:20Z</timestamp>
    <contributor>
      <username>JarBot</username>
      <id>1112010</id>
    </contributor>
    <minor/>
    <comment>بوت: إضافة قالب</comment>
    <model>wikitext</model>
    <format>text/x-wiki</format>
    <text xml:space="preserve" bytes="522">{{ {{{1&lt;noinclude&gt;|بلد بيانات عرض&lt;/noinclude&gt;}}}
 alias = الجزائر
 flag alias = Flag of Algeria.svg
 flag alias-1958 = Flag of Algeria (1958-1962).svg
 flag alias-naval = Naval Ensign of Algeria.svg
 link alias-naval = القوات البحرية الجزائرية
 size = {{{size|}}}
 name = {{{name|}}}
 altlink = {{{altlink|}}}
 variant = {{{variant|}}}
lt;noinclude&gt;
 var1 = 1958
 redir1 = DZA
 redir2 = ALG
 cat = جزائر
lt;/noinclude&gt;
}&lt;noinclude&gt;{{حماية قالب}}&lt;/noinclude&gt;</text>
    <sha1>p3nb5n16asivb42hhsyvh61w2h5qcf1</sha1>
  </revision>
</page>
```


FIGURE 2.9 – exemple code infbox de pays

• Ce modèle inclut des informations sur les pays ainsi que des pages Web externes contenant des informations détaillées sur Lod.

- **l'infobox liens de pages :**

```
|formatStatements
|property=P21
|ent{{#لو:{{{entityId|{{{{معرف ويكي بيانات}}}}}}|ityId}}={{{entityId|{{{{معرف ويكي بيانات}}}}}}
|rank=best
|formatting=raw
|noref=true}}
|Q6581097={{{1|}}}
|Q6581072={{{2|}}}
||#default={{ {{{{safesubst:}}}#1}}|{{{{|3}}}|{{{{|3}}}:لو|}}}}}
}}&lt;/includeonly&gt;&lt;noinclude&gt;
{{توثيق}}
[[تصنيف:قوالب ويكي بيانات]]&lt;/noinclude&gt;</text>
```

FIGURE 2.10 – exemple code infobox de page web

• Le modèle d'identifiant de la page web contient la qualité du site ainsi que la méthode d'accès et l'identifiant Wikipédia qui contient les informations pour cette page

2.4 Conclusion :

Nous avons présenté une explication simplifiée des caractéristiques de la Wikipédia arabe et la caractéristique de la langue arabe et fournirons une explication des exemples de modèles, y compris le modèle de l'histoire ... compris le modèle Link site WikipediaArab et la différence entre l'arabe et les autres langues Nous avons également donné des exemplesArchitecture du FrameworkDBpedia de personnages historiques algériens tels que Houari Boumediene

Chapitre 3

Proposer un algorithme d'extraction des données structurées depuis wiki arabe

3.1 Introduction :

Dans ce chapitre nous nous référerons à l'étude appliquée et aux étapes que nous avons accomplies depuis l'installation des éléments et des sabots dans l'ordinateur pour extraire les données et les fichiers de Xml, Ttl et Sola pour proposer l'algorithme d'extraction des données.

3.2 Les étapes successives pour accéder au fichier ttl

1 - Installation de JDK 1.8.0 ,JRE 1.8.0 et
Notepad++ : Cette partie est réservée à l'installation de (JDK 1.8.0, JRE 1.8.0)10 [1],eclipse11, Notepad++ 12. [2]

2 - Placez les deux dossier : extraction-framework-master ; apache-maven-3.5.3 dans d :/

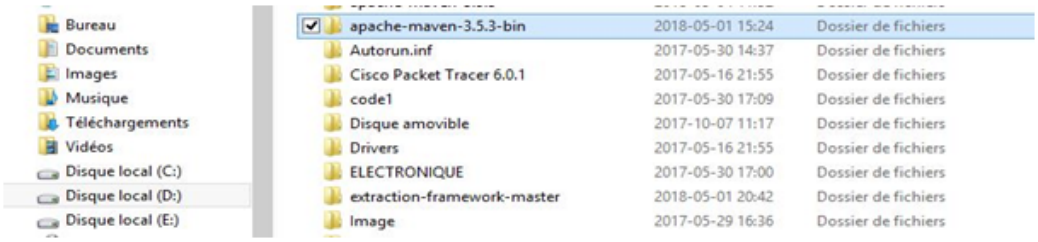


FIGURE 3.1 – Placez les deux dossier dans d :/

1. www.java.com
2. https ://notepad-plus-plus.org/download/v7.5.6.html ;

3 - créationdes variables d'environnement : La deuxième étape consiste à créer des variables. d'environement

— **a)** Nouvelle variable :JAVA_HOME.

Valeur : C :\Program Files\Java\jdk1.8.0_171\bin

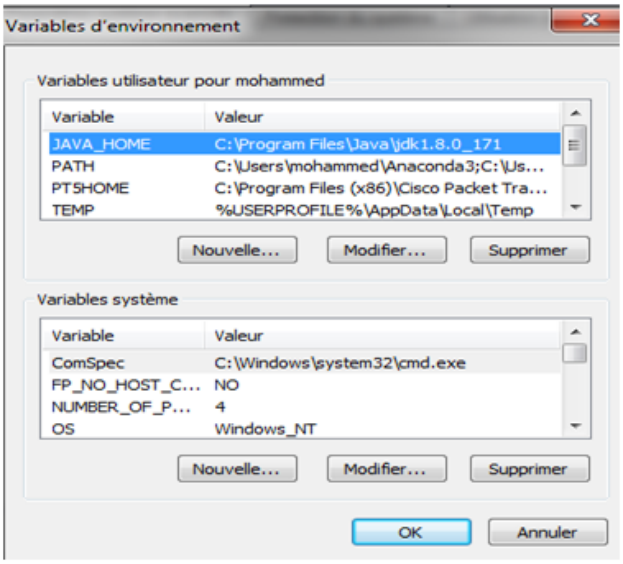


FIGURE 3.2 – créationdes variables d'environnement

b) dans la variable path ajouter les valeurs suivantes :

D :\apache-maven-3.5.3-bin \apache-maven-3.5.3\bin ;

C :\Program Files\Java\jre1.8.0_171\binModifier

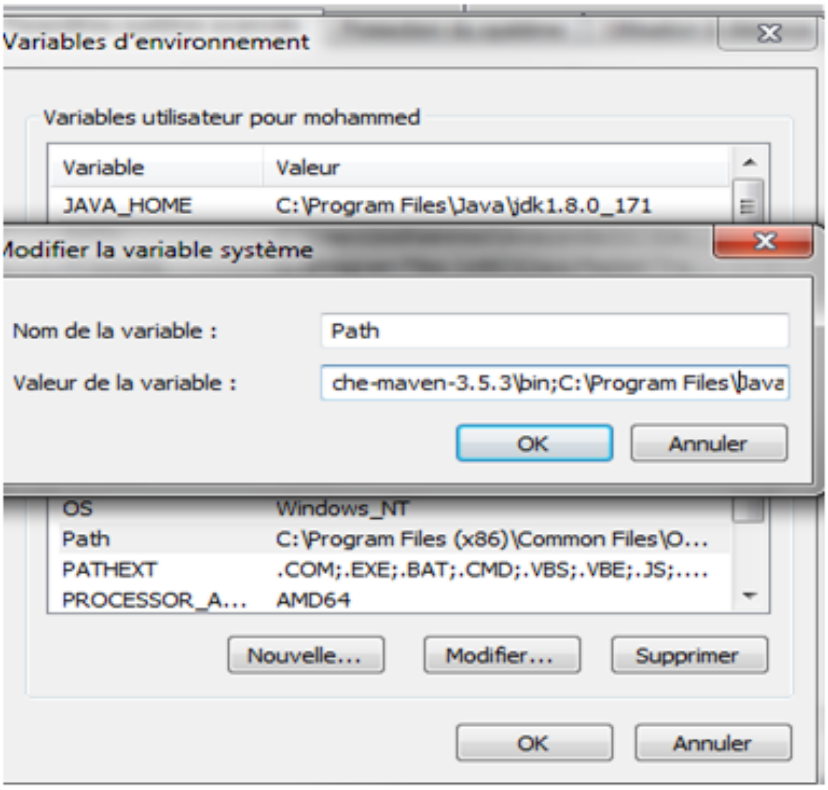


FIGURE 3.3 – ajouter les valeurs dans la variable (path)

4 - Installer maven : A l'invite commande dos en ecrire la commande suivante :

```
C:\Users\mohammed>d:

D:\>cd extraction-framework-master

D:\extraction-framework-master>cd dump

D:\extraction-framework-master\dump>mvn install
```

La fenêtre suivante est apparait a la fin d'installation maven

```
[INFO]
[INFO] --- maven-install-plugin:2.4:install (default-install) @ wiktionary ---
[INFO] Installing D:\extraction-framework-master\wiktionary\target\wiktionary-4.
2-SNAPSHOT.jar to C:\Users\mohammed\.m2\repository\org\dbpedia\extraction\wiktio
nary\4.2-SNAPSHOT\wiktionary-4.2-SNAPSHOT.jar
[INFO] Installing D:\extraction-framework-master\wiktionary\pom.xml to C:\Users\
mohammed\.m2\repository\org\dbpedia\extraction\wiktionary\4.2-SNAPSHOT\wiktionar
y-4.2-SNAPSHOT.pom
[INFO] Installing D:\extraction-framework-master\wiktionary\target\wiktionary-4.
2-SNAPSHOT-javadoc.jar to C:\Users\mohammed\.m2\repository\org\dbpedia\extractio
n\wiktionary\4.2-SNAPSHOT\wiktionary-4.2-SNAPSHOT-javadoc.jar
[INFO] Installing D:\extraction-framework-master\wiktionary\target\wiktionary-4.
2-SNAPSHOT-sources.jar to C:\Users\mohammed\.m2\repository\org\dbpedia\extractio
n\wiktionary\4.2-SNAPSHOT\wiktionary-4.2-SNAPSHOT-sources.jar
[INFO] Installing D:\extraction-framework-master\wiktionary\target\wiktionary-4.
2-SNAPSHOT-jar-with-dependencies.jar to C:\Users\mohammed\.m2\repository\org\dbp
edia\extraction\wiktionary\4.2-SNAPSHOT\wiktionary-4.2-SNAPSHOT-jar-with-depende
ncies.jar
[INFO] ------------------------------------------------------------------------
[INFO] Reactor Summary:
[INFO]
[INFO] Parent POM of the DBpedia framework 4.2-SNAPSHOT ... SUCCESS [03:45 min]
[INFO] DBpedia Core Libraries ............................. SUCCESS [10:45 min]
[INFO] DBpedia Server ..................................... SUCCESS [01:11 min]
[INFO] DBpedia Dump Extraction ............................ SUCCESS [ 30.395 s]
[INFO] DBpedia Scripts .................................... SUCCESS [01:13 min]
[INFO] Live extraction .................................... SUCCESS [01:24 min]
[INFO] Wiktionary Dump 4.2-SNAPSHOT ....................... SUCCESS [01:36 min]
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time: 21:51 min
[INFO] Finished at: 2018-05-14T21:02:09+01:00
[INFO] ------------------------------------------------------------------------

D:\extraction-framework-master>
```

5 - Télécharger ficher wikipedia arabe (*.xml) : Pour obtenir un fichier (*.xml), nous allons sur la page web :

https ://ar.wikipedia.org/wiki/%D8%AE%D8%A7%D8%B5 : %D8%AA%D8%B5%D8%AF%D9%8A%D8%B1 image suivant :

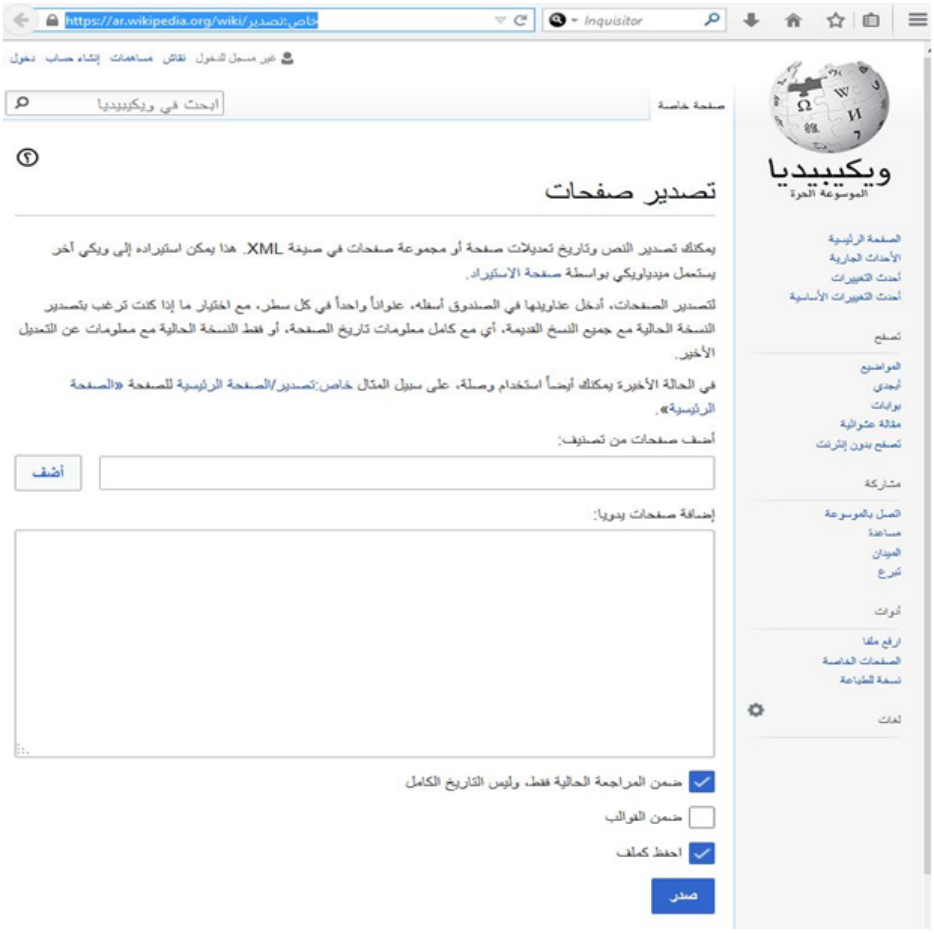
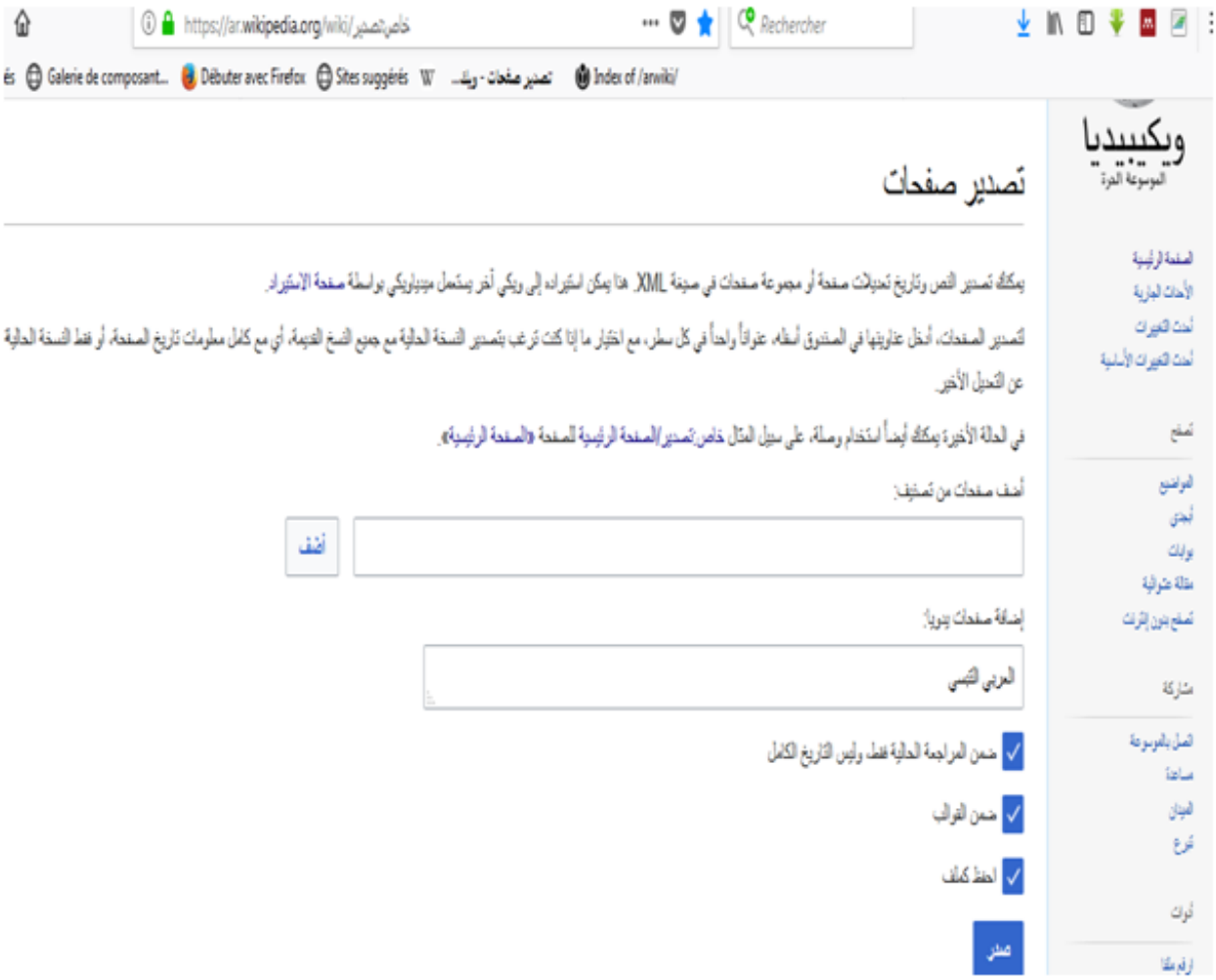


FIGURE 3.4 – site d'exportation wikipédia arabe

- Entrez la phrase que vous voulez rechercher* .xmlEn إضافة صفحات يدويا par ex : العربي التبسي
- Et cochez la case : ضمن القوالب
- Cliquez sur : صدر



La **fenêtre** suivante est apparait.

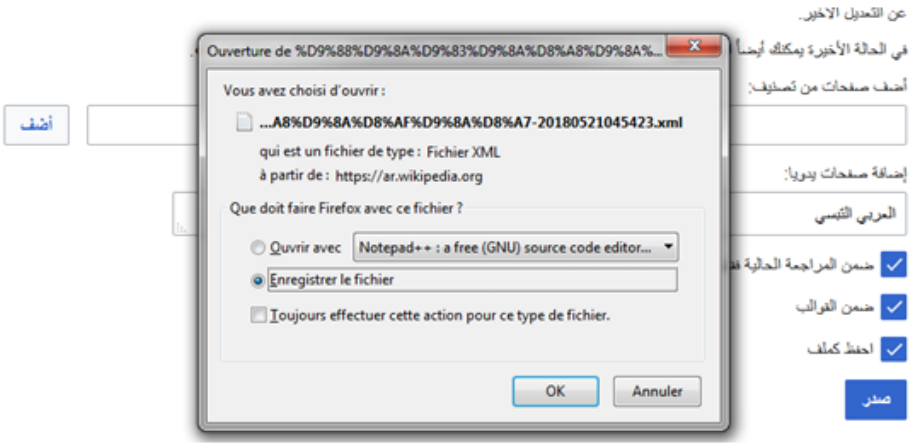


FIGURE 3.5 – enregestré page xml à partir le site

Cliquer sur ok le fichier xml est télécharger dans le repertoiredownload

Rennomercet fichier comme suit : arwiki-20180301-pages-articles.xml

Déplacerárwiki-20180301-pages-articles.xmláu :D : \extraction-framework-master \dump\ala\arwiki\20180301

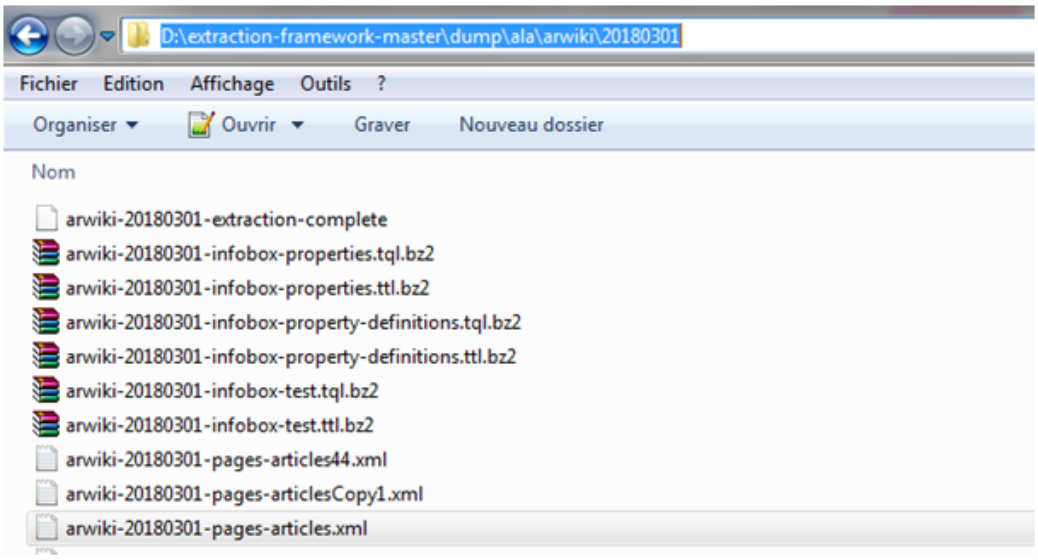


FIGURE 3.6 – Lorsque nous ouvrons le fichier par nodepad++ , nous trouvons ce qui suit

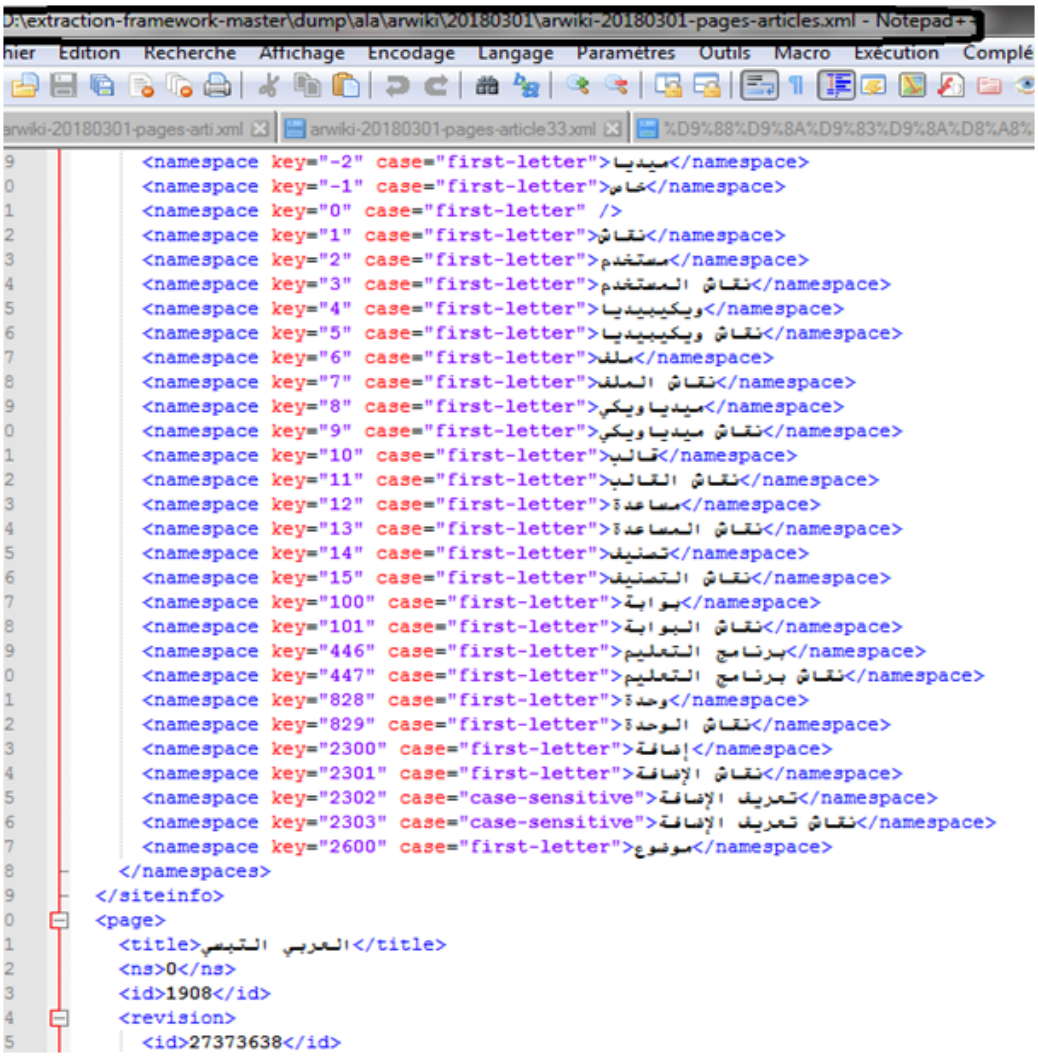


FIGURE 3.7 – Fichier xml aprés l'exportation

3.3 extraction des données à partir fichier xml :

1. aller vers D :\extraction-framework-master\dump.

2. ouvrir le fichier extraction.default par nodepad++ et ensuite vérifier

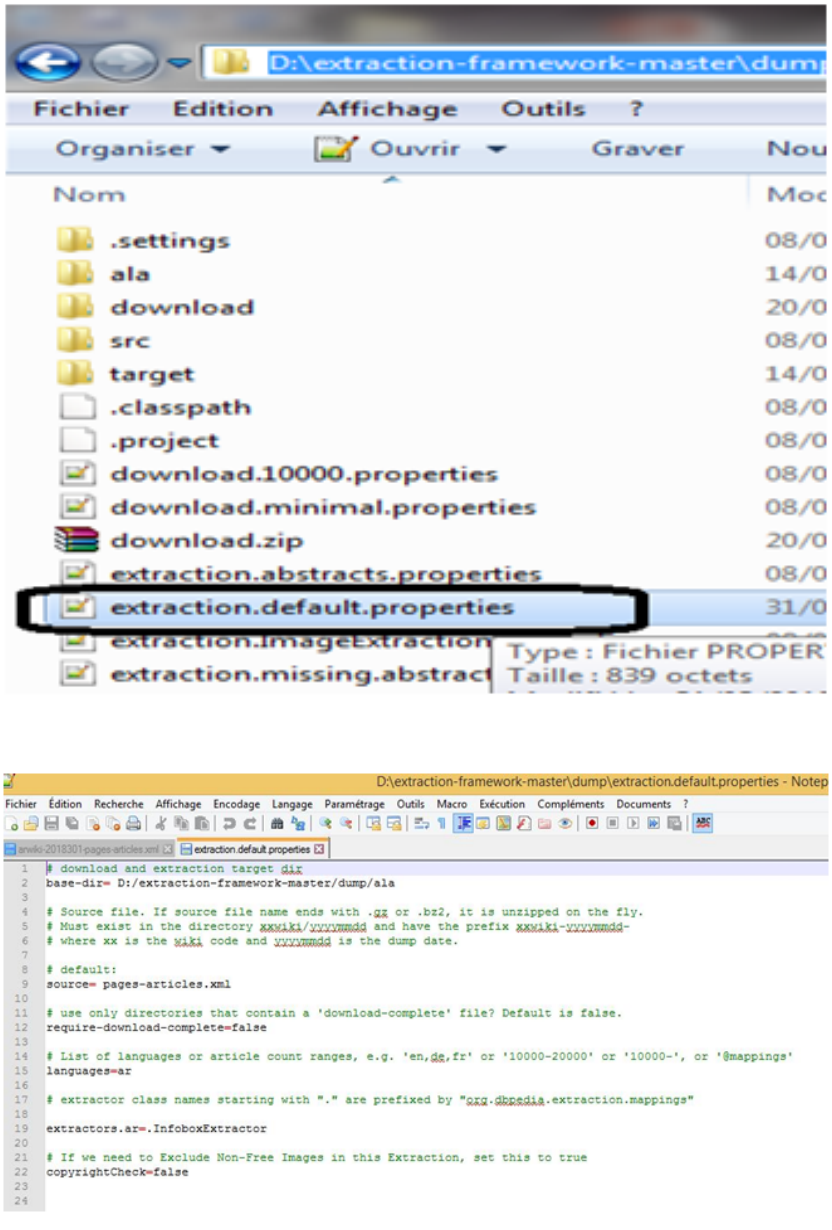


FIGURE 3.8 – Extraction default properties

3. Vérifier que le fichier contient les lignes suivantes : base-dir= D :/extraction-framework-master/dump/ala
source= pages-articles.xml
require-download-complete=false
languages=ar
extractors.ar=.InfoboxExtractor
copyrightCheck=false

4. ouvrir le mode commande DOS dans le dossier dump.

5. taper la commande suivante pour l'extraction :

mvnscala :run -Dlauncher=extraction -DaddArgs=extraction.default.properties

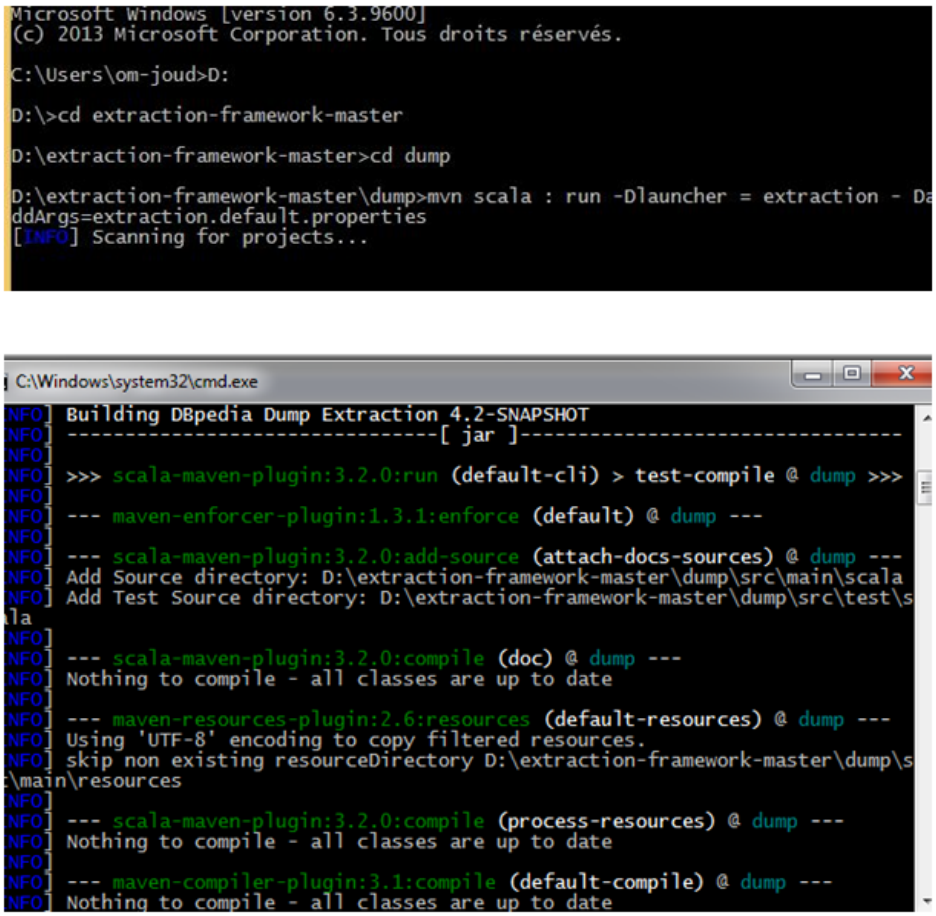


FIGURE 3.9 – extraction les fichiers ttl et tql

Le résultat est dans le dossier : D :\extraction-framework-master\dump\ala\ arwiki\20180301 :

arwiki-20180301-infobox-properties.ttl et arwiki-20180301-infobox-properties.tql.

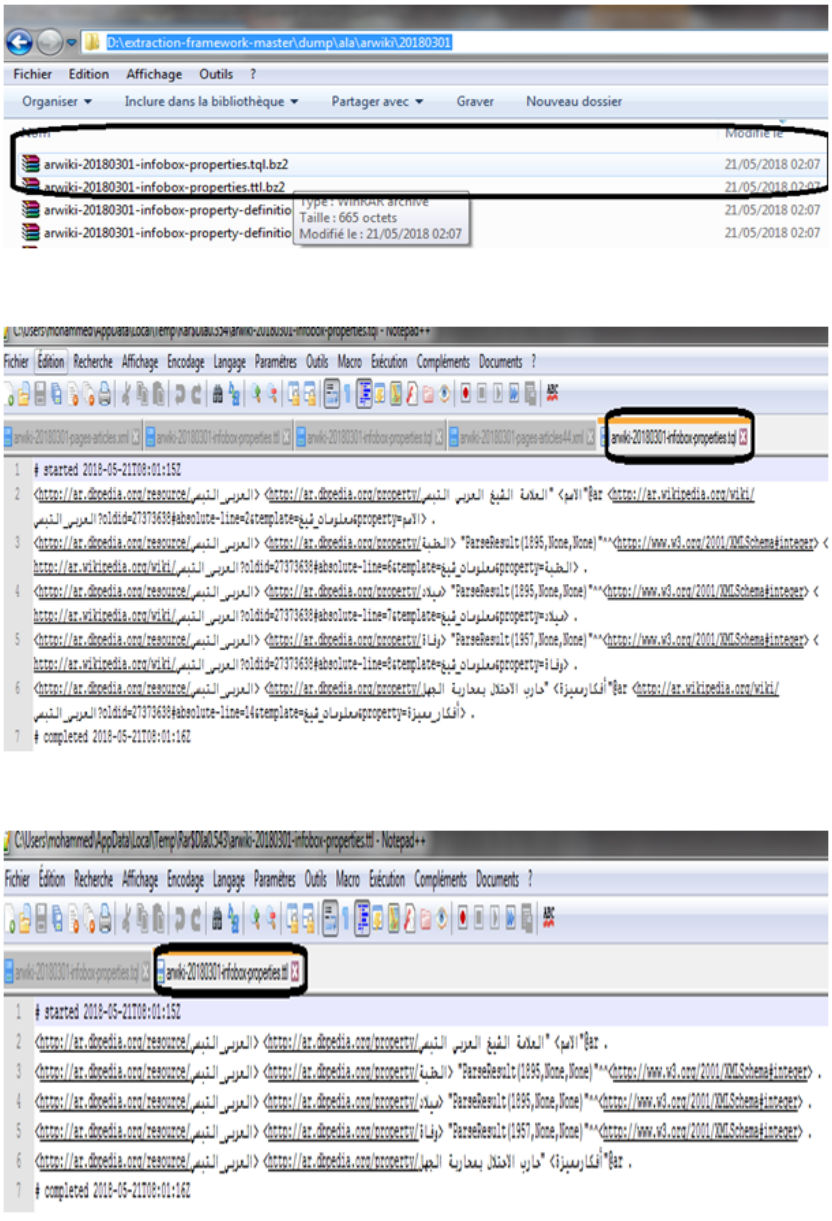


FIGURE 3.10 – Contenu de fichier ttl

Autre exemple pour résultat ttl(Mohammed Chebaiki : Poète de la révolution algérienne) :

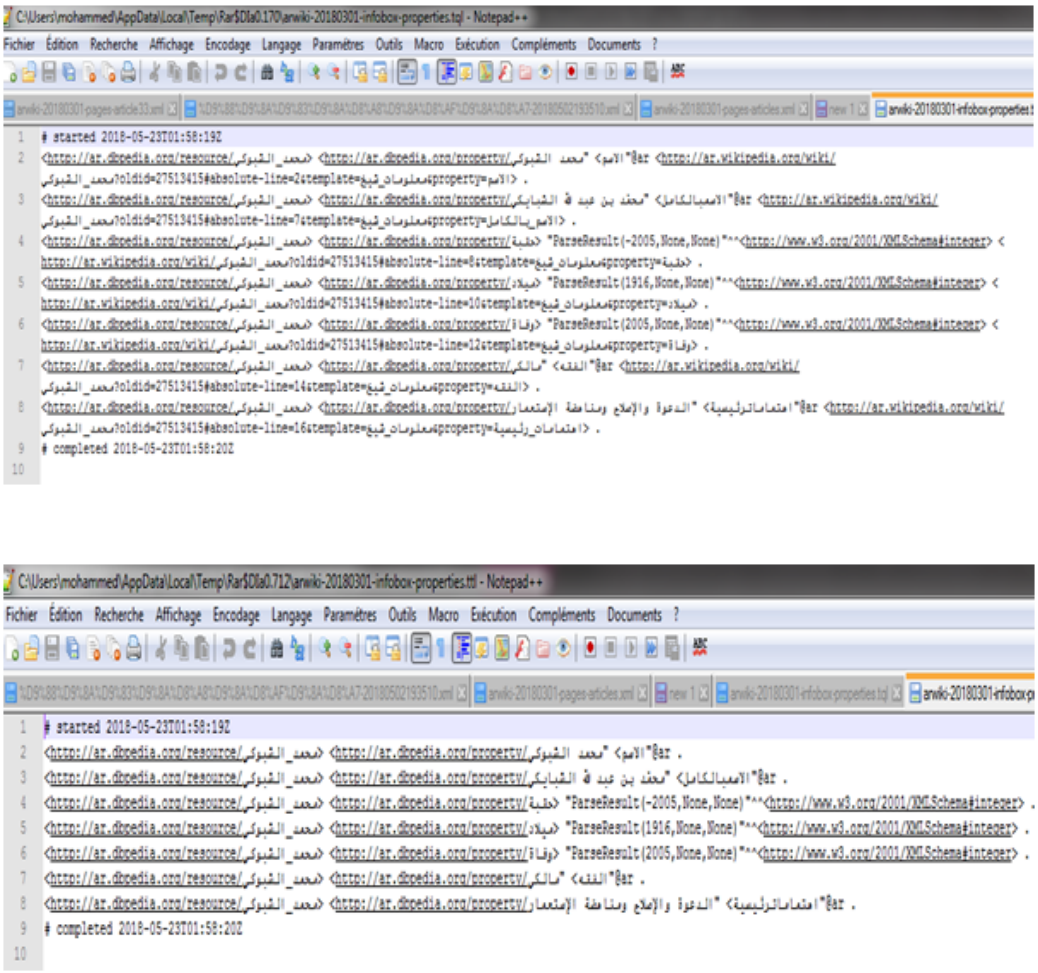


FIGURE 3.11 – contenu de fichier ttl (moète Mohammed Chebaiki)

3.4 Créer un algorithme d'extraction :

À ce stade, nous allons créer un algorithme d'exportation que nous utiliserons plus tard dans la création d'un programme qui convertit le fichier *.XML en un fichier *.TTL.

3.4.1 Le diagramme de l'algorithme :

Le diagramme suivant illustre les phases et les sections de l'algorithme

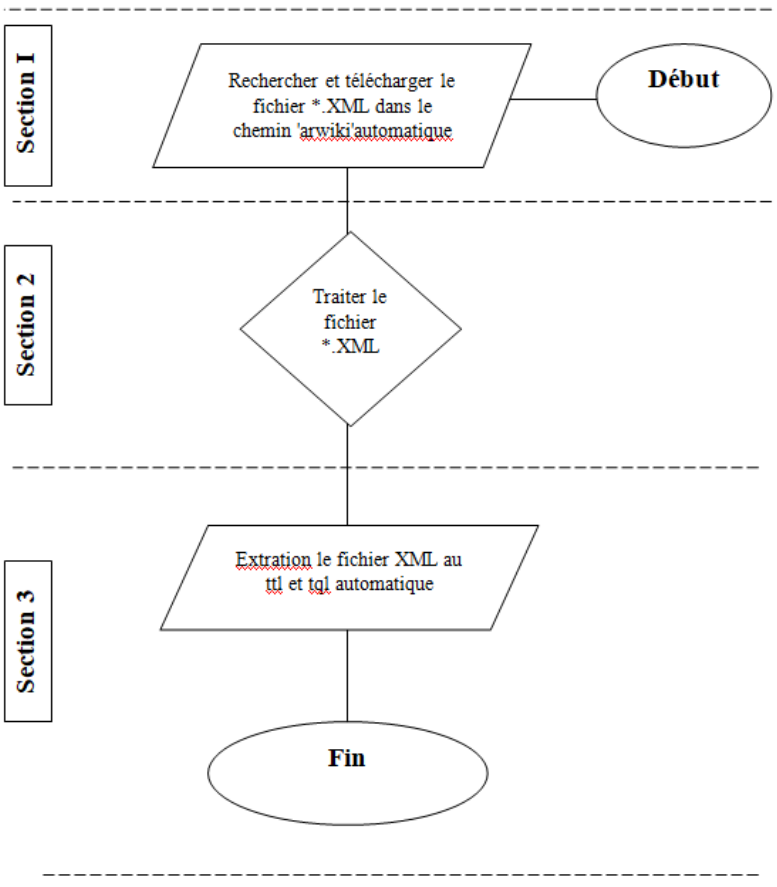
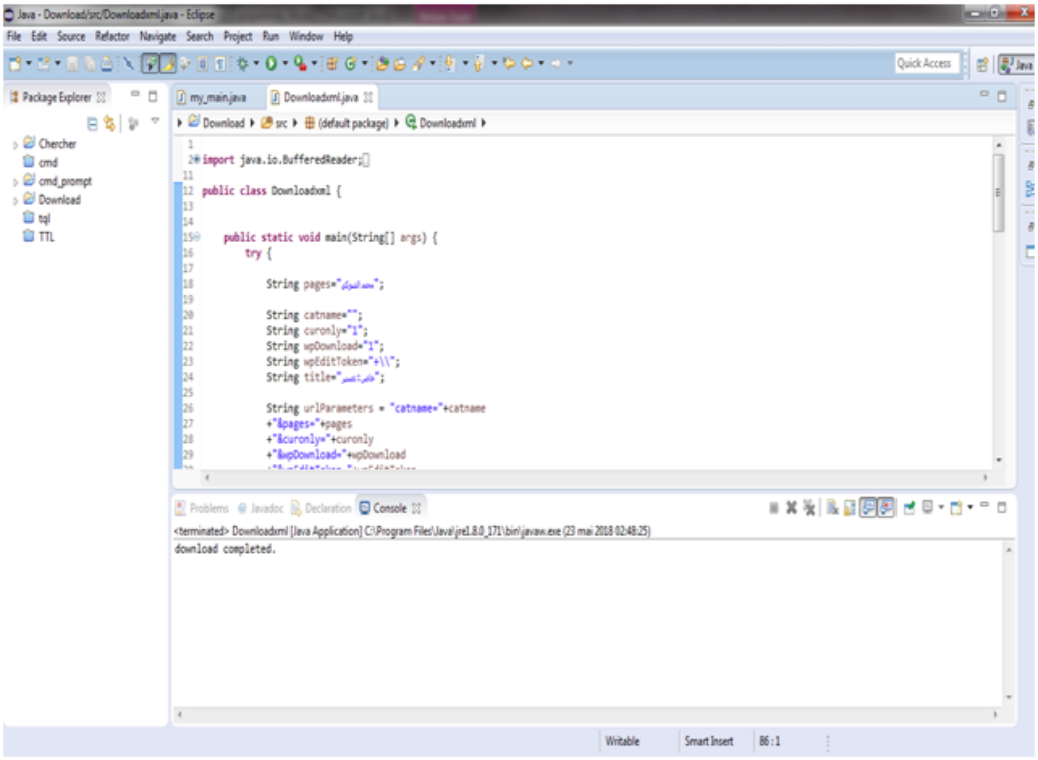


FIGURE 3.12 – Le diagramme de l'algorithme

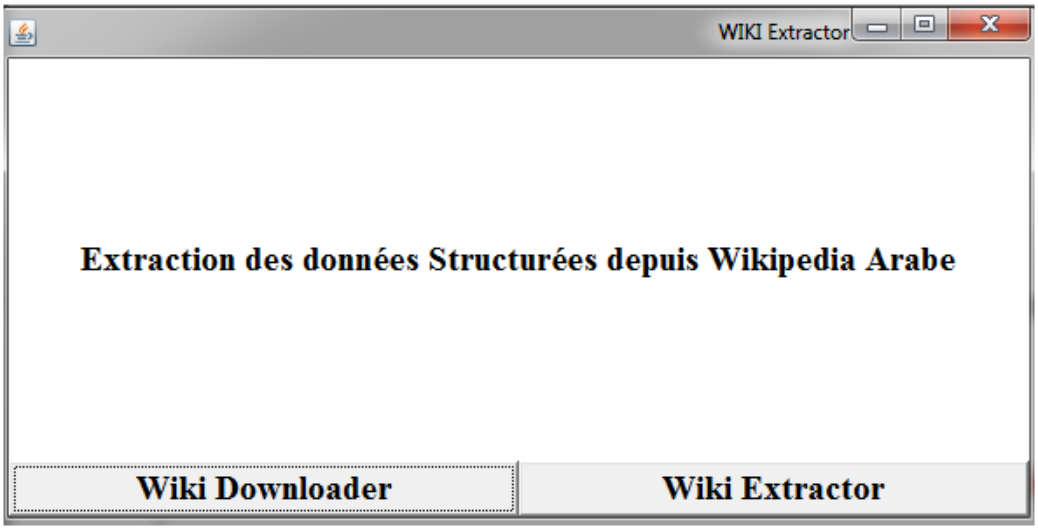
3.4.2 Le code avec JAVA :

A) - phase une :




```
1
2    String pages="محمد الشبوكي";
3    String title="خاص:تصدير";
4    String urlParameters = "catname="+catname;
5    String request = "https://ar.wikipedia.org/wiki/%D8%AE%D8%A7%D8%B5:%D8%AA%D8%B5%D8%AF%D9%8A%D8%B1";
6            {
7          response.append(inputLine);
8          response.append("\n");
9                }
10       in.close();
11            String filePath="D:/extraction-framework-master/dump/ala/arwiki/20180301/";
12           String fileName="arwiki-20180301-pages-articles.xml";
13         System.out.println(ex.getMessage());
14   }
15
```

FIGURE 3.13 – Code java de téléchcharger fichier xml



- cliquez sur le bouton " Wiki Downloader",la fenêtre suivant apparait.

Entrez la phrase que vous voulez rechercher,et cliquez sur le boutton "Download"

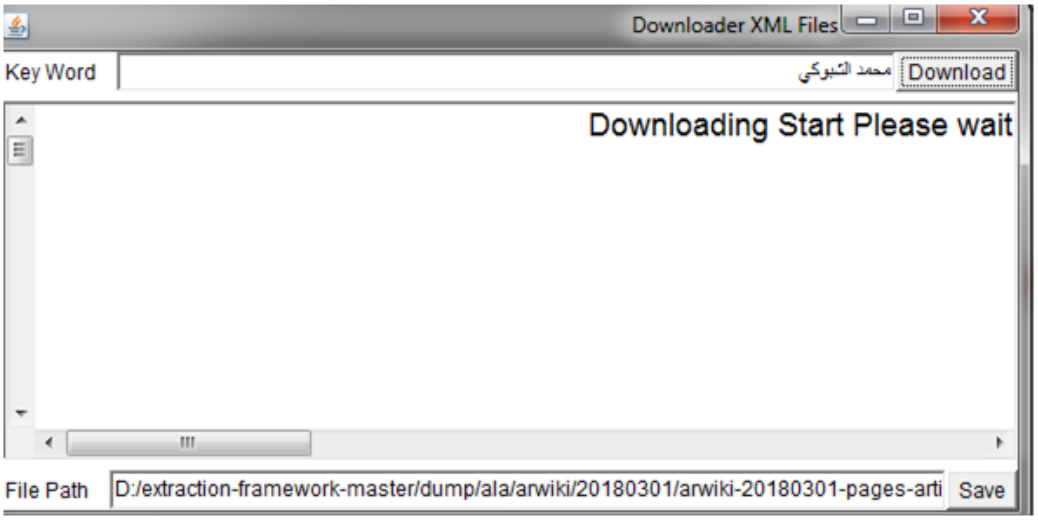


FIGURE 3.14 – télécharger le fichier xml

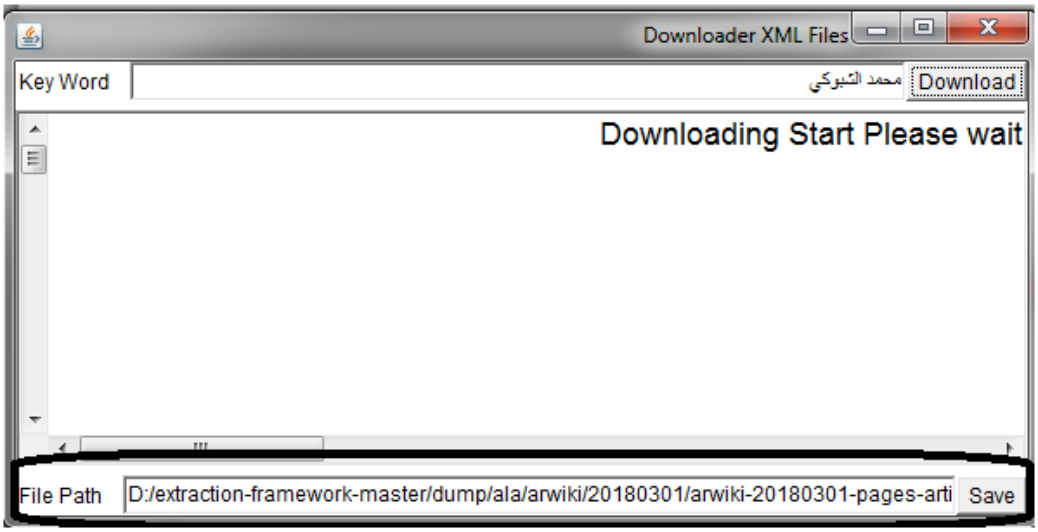
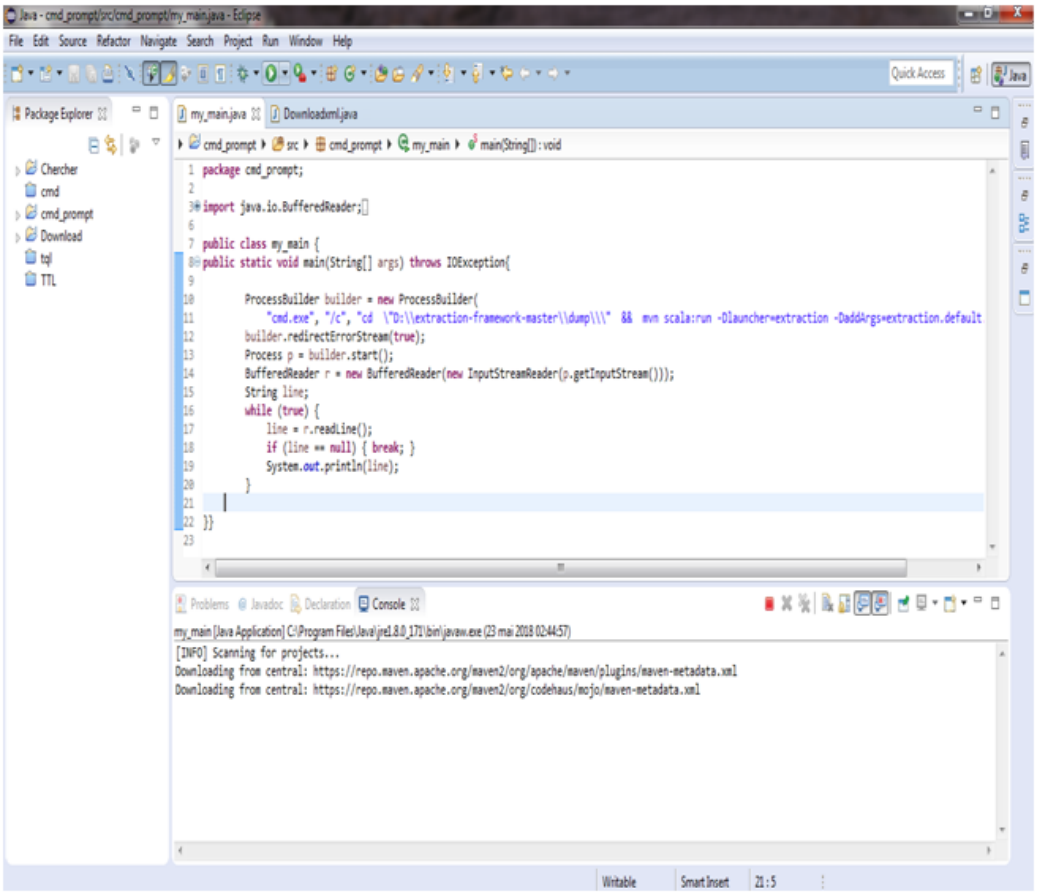


FIGURE 3.15 – Enregestré le fichier xml dans le chemin :... \20180301

B) - phase deux : extraction des données structurées a partir fichichier XML et le résultat est dans meme chemin (D :\extraction-framework-master\dump\ala\arwiki\20180301)




```
1
2   ProcessBuilder builder = new ProcessBuilder("cmd.exe", "/c", "cd  \"D:\\extraction-framework-master\\dump\\\""
3                   &&mvnscala:run-Dlauncher=extraction-DaddArgs=extraction.default.properties");
4   builder.redirectErrorStream(true);
5   }
```

FIGURE 3.16 – code java pour d'extraction le fichier xml

- cliquez sur le boutton "extraction" les fichier : arwiki-20180301-infobox-properties.ttl et arwiki-20180301-infobox-properties.tql comme une résultat dans : (D :\extraction-framework-master\dump\ala\arwiki\20180301).

- quand on presse le boutton"Openttl" les fichiers : arwiki-20180301-infobox-properties.ttl et arwiki-20180301-infobox-properties.tql sont apparait.

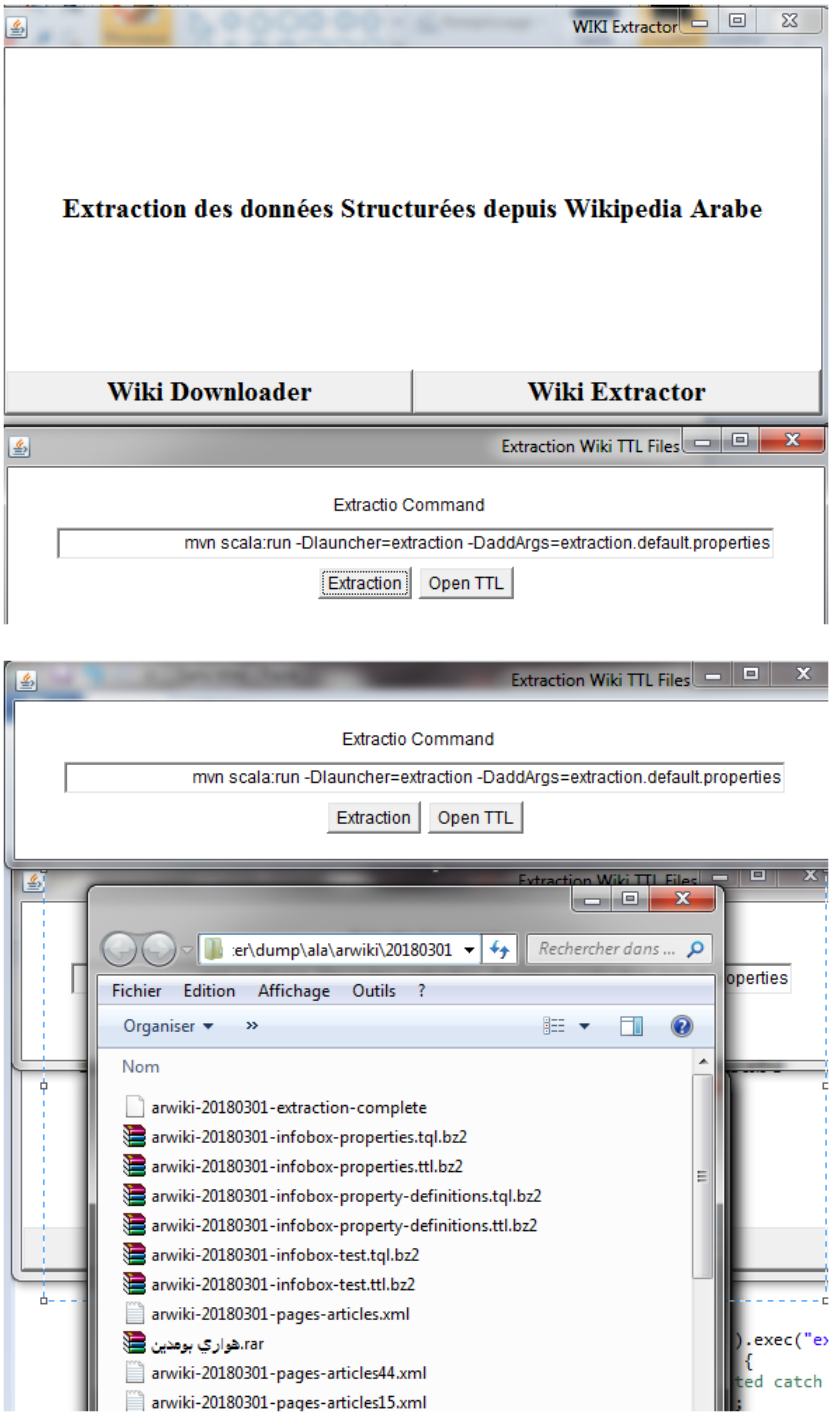


FIGURE 3.17 – l'emplacement de fichier ttl

3.5 Conclusion :

Dans ce chapitre, nous avons introduit un algorithme qui permet de télécharger un fichier Wikipedia en arabe sous la forme d'un fichier xml dans la première phase, puis de le convertir en données structurées sous la forme d'un fichier ttlauphase deuxième.

Conclusion Générale :

Conclusion et suggestions :

Dans cette mémoire, nous pouvons dire que c'est le début que nous avons été en mesure de créer un algorithme qui permet le téléchargement et d'extraction depuis Wikipédia arabe dans un sujet donné et par conséquent nous donne des données structurées sous la forme de RDF.

Malgré le manque d'information, l'extraction des données structurées depuis Wikipédia arabemais malheureusement l'arabe n'est pas totalement disponible, contrairement à d'autres langues,

Nous n'avons utilisé qu'une extraction.default.properties et comme étude future nous pouvons généraliser cet algorithme à différents extracteurs (extracteurs différents)

Comme suggestions futures, nous aimerions mentionner ce qui suit :

— Algorithme amélioré et expansion Il en a fait une application efficace sur Internet pour Wikipédia arabe.
— Créer un lien entre les différentes universités arabes pour enrichir ce sujet important et la recherche et le développement à travers les forums et forums ....
— Établir des bases de données pour la conservation et l'extraction de données organisées et au niveau des serveurs.

Bibliographie

[1] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann :, *"DBpedia-A Crystallization Point for the Web of Data"* Journal of Web Semantics, vol. 7, pp. 154-165, September 2009.

[2] Wikipedia [Online].Available : http ://en.wikipedia.org/wiki/University.[Accessed : September. 6, 2013].

[3] World Wide Web Consortium.[Online].
Available : http ://www.w3.org/2009/sparql/wiki/Main_Page[Access ed : September. 5, 2013].

[4] D. Lange, C. Böhm, and F. Naumann, "Extracting Structured Information from Wikipedia Articles to Populate Infoboxes," *in The 19th ACM international conference on Information and knowledge management*, pp. 1661-1664.

[5] Haytham Al-Feel, *Ph.D,International Journal of Computer Applications (0975 – 8887)* Volume 80 – No3, October 2013

[6] Sebastian Hellmann, *Claus Stadler*, et al. (2009). DBpedia Live Extraction OTM *'09 Proceedings of the Confederated International Conferences, Springer- Verlag Berlin*, Heidelberg. pp. 1209 / 1223

[7] Pablo N. Mendes, Max Jakob, et al. (Sept. 7-9, 2011). DBpedia Spotlight : *Shedding Light on the Web of Documents. I-SEMANTICS* , 7th Int. Conf. on Semantic Systems, *Graz, Austria.*

[8] M.Morsey, S. Auer, J. Lehmann,C.Stadler and S.Hellmann, *"Dbpedia and the Live Extraction of Structured Data from Wikipedia,"*In Pro-

gram :electronic library and information systems,*Vol 46(2012)*,p.27.

[9] Haytham Al-Feel, Ph.D,International Journal of Computer Applications (0975 – 8887) Volume 80 – No3, October 2013

[10] http ://prefix.cc provides list of commun prefixes