

# Chapitre 1

## Contribution

### 1.1 Ouverture

Les articles de Wikipédia sont constitués généralement d'un texte, mais aussi de certaines informations structurées (infobox, images, liens externes, redirections entre les pages etc...) présentes sous la forme de balises Wiki. Le projet DBpedia extrait des informations structurées de Wikipédia et les transforme dans une base de connaissances riche sous forme d'un graphe avec des entités reliées.

Dans ce chapitre, nous allons vous donner une vue d'ensemble sur l'annotation des métadonnées, une analyse des besoins, la procédure d'extraction de DBpedia et les pistes de travail possibles. Puis nous présenterons l'architecture globale de notre système ainsi que notre hypothèse et la structure de l'application que nous avons développée pour concrétiser notre hypothèse. La dernière partie de cette section porte sur les résultats de cette étude.

### 1.2 Utilité des annotations

En générale, l'annotation, c'est une étiquette qu'on ajoute à une ressource Web. Depuis la création du Web, plusieurs systèmes d'annotation sont apparus (ThirdVoice, PageSeeder, HyperNews, Nestor, etc...). Nous citons brièvement les conséquences liées à ces systèmes d'annotation telles que l'information annotée doit d'une manière ou d'une autre être structurée, utilisable et descriptive de la ressource ou de son utilisation. De plus, la ressource en

question doit exister et peut être exploitée sur le Web indépendamment des informations qui lui sont associées. La figure ci-dessous montre le système intermédiaire entre le client et le service Web dans lequel il y a le service de gestion des annotations, permettant la communication entre ces deux entités.

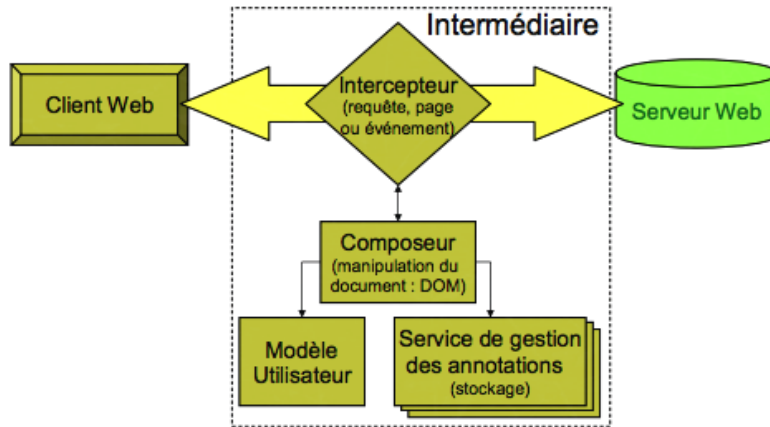


FIGURE 1.1 – Notation d'intermédiaire

L'annotation sémantique fait référence à plusieurs types distincts d'annotations formelles, explicites et permanentes. Il existe des outils d'annotation basés sur les ontologies *Ontology based annotation tool* et des critères relatifs aux annotations par exemple : les types de ressources concernées, la structuration des schémas de description, l'automatisation marquée de la mise en place, etc.

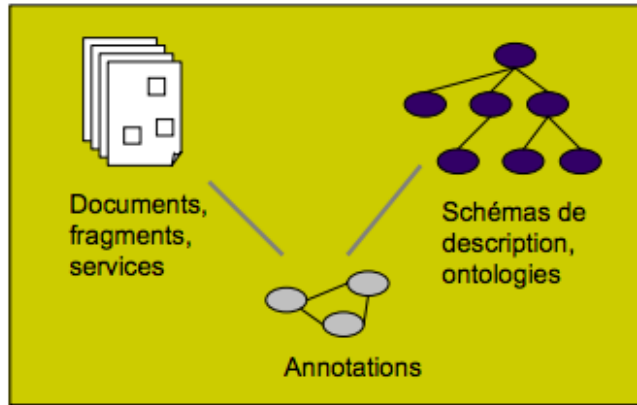


FIGURE 1.2 – Différents niveaux de connaissances

L’annotation d’un triplet RDF est une façon d’ajouter des metadonnées à un triplet RDF pour décrire une restriction spatiale.

*Comment on utilise les annotations temporelles ?* Un exemple d’utilisation est sur le site “sig.ma”<sup>1</sup> créée par *the digital enterprise research institute in Ireland*, la plateforme fournit un moteur de recherche par mot clé qui permet de récupérer des images et des textes accessibles par des annotations RDF, ainsi qu’une liste d’URI synonymes correspondant à la clé de recherche et des liens vers des sources Web contenant des données RDF pertinentes.

### 1.3 Analyse des Besoins

Le large succès de Wikipedia (qui est le 2ème site le plus visité sur internet) et le progrès des techniques d’extraction des données ont abouti à la naissance de la construction automatique de larges bases de connaissances comme DBpedia, YAGO, etc...

---

1. <http://sig.ma/>



FIGURE 1.3 – Différentes sources d'informations dans DBpedia

Beaucoup de connaissances sont construites en se basant sur l'extraction automatique des faits relationnels dans un texte. En effet, les bases de connaissances convergent sur les faits statiques et ne donnent pas une grande importance à la dimension temporelle de ces triplets. Et ceci a lieu en dépit du fait que la majorité des faits évoluent avec le temps, ou ne sont valides que dans une période temporelle précise. Ainsi, nous remarquons que le temps a une dimension significative dans ces bases de connaissances.

La dimension temporelle est particulièrement importante dans les relations binaires comme *isPresidentOf*, *isCEOOf*, *isMarriedTo*, on peut être mariée à plusieurs épouses mais dans des différents intervalles de temps (On ne tiens pas compte des exceptions que représentent les mariages polygames). Une base de connaissances contenant plusieurs présidents des États-Unis ne peut être consistante que lorsqu'on ajoute une dimension temporelle à ces faits. De plus l'annotation temporelle aide à faire la distinction entre les faits courants et les faits dépassés. Par exemple le fait "Kennedy est le président des États-Unis" est correct, mais n'est plus valide. Lorsqu'on attache une annotation temporelle à un fait comme celui là, il devient universellement valide.

## 1.4 Problématique

Lorsqu'on parcourt DBpedia, on trouve beaucoup de triplets qui décrivent des informations temporelles. Ces derniers sont généralement liés à un contexte événementiel précis. Il est plus difficile d'exploiter ces informations si elles ne possèdent pas une structure universellement valide, claire et lisible par la machine. Dans DBpedia, il se trouve que des informations liées au même contexte temporel sont exprimées de la manière suivante :

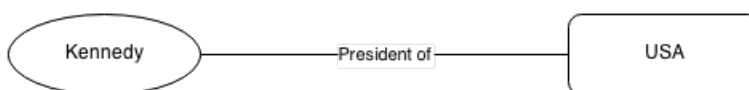


FIGURE 1.4 – triplet "Kennedy"

Le premier triplet n'a pas une sémantique valide que en tenant compte du triplet suivant :

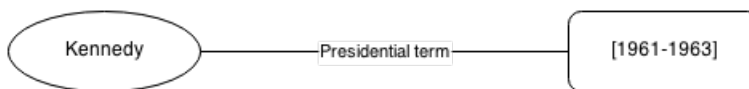


FIGURE 1.5 – triplet presidential term "Kennedy"

Dans cette étude, on vise plutôt à annoter les triplets  $(s, p, o)$  avec une étiquette temporelle qui indique et précise la validité de ce terme dans un cadre logique qui appartient au monde réel où en dehors de ce cadre, on peut dire que ce triplet RDF n'est pas valide et qu'on ne peut pas l'utiliser.

## 1.5 Étude préliminaire et approches possibles

### 1.5.1 Web Collaboratif

C'est le Web qui s'appuie sur les utilisateurs pour construire son contenu. Nous avons commencé notre travail de recherche par une étude préliminaire autour du contenu de ces plateformes collaboratives. Aussi, nous avons étudié les pistes possibles pour l'exploitation des dumps de Wikipédia et Wikidata. Tout d'abord, nous avons téléchargé les fichiers des collections XML et nous avons observé la structure des informations dans ces sources d'informations.

Ensuite nous avons implémenté un premier algorithme d'extraction en utilisant un parseur XML (SAX<sup>2</sup>). La figure ci-dessous représente notre schéma de modélisation dans lequel nous avons procédé avec une modélisation qui touche directement la source principale d'informations Wikipédia.

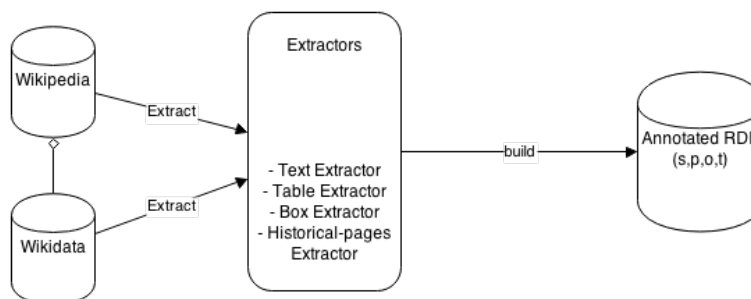


FIGURE 1.6 – Première approche : schéma de modélisation générale

Cette modélisation est une première rubrique d'analyse et de conception d'une solution qui touche les besoins préliminaires de notre étude. Par ailleurs, nous avons retrouvé une autre modélisation plus proche à nos besoins principaux et que nous vous détaillerons par la suite.

### 1.5.2 Traitement automatique des langues

Le traitement automatique de la langue (TAL) est une discipline à la frontière de la linguistique, qui est intimement liée à l'intelligence artificielle. Il existait un type de TAL statistique proposant des méthodes statistiques, probabilistes ou purement statistiques pour résoudre certaines difficultés. On distingue plusieurs domaines d'application de TAL comme la traduction automatique, la génération automatique de texte, la correction orthographique, la reconnaissance de l'écriture manuscrite, etc... Mais dans ce type de traitement, il y a des problèmes qui peuvent apparaître et principalement celui de l'ambiguïté temporelle.

### 1.5.3 Ambiguïtés temporelles

C'est la propriété d'un mot ou d'une suite de mots (comme notre cas) qui peuvent avoir un ou plusieurs sens d'analyses grammaticales possibles. Dans

2. [http://fr.wikipedia.org/wiki/Simple\\_API\\_for\\_XML](http://fr.wikipedia.org/wiki/Simple_API_for_XML)

une phrase simple ou composée, l'indicateur temporel peut avoir plusieurs sens tout dépend du contexte de la phrase.

Les informations temporelles peuvent avoir des représentations différentes :

- Un évènement “Je vous propose un rendez-vous *demain* pour parler de ma plateforme PiSharing”.
- Une connaissance “Jacques Chirac est le président de la république Française” **mais quand ?**.

Le présent par exemple peut avoir plusieurs sens ou contextes : présent de narration, présent de généralité, présent qui réfère au futur proche, etc... Les signaux temporels sont ambigus par exemple dans ces expressions : il court pour rattraper le temps, tu tournes après la rivière, etc. . . On remarque qu'il y a des indicateurs temporels, mais ce n'est pas le temps qui est relatif à un événement qui peut nous intéresser. La plupart des expressions sont floues comme : il y a deux ans, chaque deux semaines, j'arrive dans deux secondes, etc... En effet, il n'y a pas une logique descriptive qui peut nous aider à mettre un lien entre l'événement et la période temporelle.

L'analyse du temps s'inscrit dans la compréhension globale des textes, et des événements auxquels on fait référence dans ce texte non pas en analysant une phrase comme suit.

Modalité : “l'équipe de France voulait gagner la coupe du monde en 2006.”

Anaphore : “..., cela pourrait avoir lieu dans les éditions suivantes.”

Les événements décrits (et que l'on souhaite fixer temporellement) peuvent être : duratifs ou ponctuels/accomplis ou inaccomplis. De même pour les dates qui peuvent être des dates absolues “le 18 mars, c'est mon anniversaire”; ou bien des dates relatives par rapport au moment de l'énonciation par exemple : “il y a deux ans”. Pour la durée aussi on distingue plusieurs types comme la durée absolue “durant 2 ans” et la durée relative “depuis un an”. Dans un texte, on trouve aussi un ensemble d'expressions de fréquence comme “tous les ans, le vendredi 13” et des expressions plus complexes comme “après la Révolution Tunisienne”.

Les textes contiennent des informations temporelles de taille massive qui sont difficilement exploitables. Nous avons donné une vue globale sur cette procédure que nous avons décidé de ne pas l'adopter parce que notre

#### 1.5.4 Historique des modifications dans Wikipédia

# Historique des versions de « Moyen Âge »

Voir les opérations sur cette page

Naviguer dans l'historique

À partir de l'année (et précédentes) :  À partir du mois (et précédents) : tous  Filtrer les balises :  ☐ Masqués seulement

Outils externes et statistiques

Liste des auteurs - Rechercher l'auteur d'un passage de l'article - Modifications - Consultations - Qui suit cette page ?

Autres discussions [liste]

Suppression - Neutralité - Droit d'auteur - Article de qualité - Bon article - Lumière sur - À faire - Archives - Traduction

Légende : (actu) = différence avec la version actuelle - (diff) = différence avec la version précédente - m = modification mineure

(dernière page | première page) Voir (50 plus récentes | 50 plus anciennes) (20 | 50 | 100 | 250 | 500).

☒ Avertir le contributeur de la demande de purge d'historique — Source copiée :

- (actu | diff) ☐ ☒ 6 juin 2011 à 18:51 Althiphika (discuter | contributions) (+2919) (LiveRC : Révocation des modifications de 82.127.154.70 (retour à la dernière version de Salebot)) (défaite)
- (actu | diff) ☐ ☒ 6 juin 2011 à 18:50 82.127.154.70 (discuter) (-2919) (→Définition) (défaite) (Balise : longue chaîne de caractères sans espace)
- (actu | diff) ☐ ☒ 6 juin 2011 à 18:48 Salebot (discuter | contributions) (+2919) (bot : révocation de 82.127.154.70 (modification suspecte : -71), retour à la version 66046433 de JLM) (défaite)
- (actu | diff) ☐ ☒ 6 juin 2011 à 18:48 82.127.154.70 (discuter) (-2919) (→Définition) (défaite) (Balise : longue chaîne de caractères sans espace)
- (actu | diff) ☐ ☒ 4 juin 2011 à 17:09 JLM (discuter | contributions) m (-79) (Annulation des modifications 66046419 de 96.23.37.85 (diff) (défaite))
- (actu | diff) ☐ ☒ 4 juin 2011 à 17:08 96.23.37.85 (discuter) (+79) (→Religion catholique) (défaite)
- (actu | diff) ☐ ☒ 31 mai 2011 à 20:04 Salebot (discuter | contributions) (-40) (bot : révocation de 216.73.72.120 (modification suspecte : -118), retour à la version 65919078 de Suprememangaka) (défaite)
- (actu | diff) ☐ ☒ 31 mai 2011 à 20:04 216.73.72.120 (discuter) (+40) (→Définition de l'Occident médiéval) (défaite)
- (actu | diff) ☐ ☒ 31 mai 2011 à 20:03 Suprememangaka (discuter | contributions) m (-24) (LiveRC : Révocation des modifications de 216.73.72.120 (retour à la dernière version de Salebot)) (défaite)
- (actu | diff) ☐ ☒ 31 mai 2011 à 20:02 216.73.72.120 (discuter) (+24) (→Définition de l'Occident médiéval) (défaite)
- (actu | diff) ☐ ☒ 31 mai 2011 à 20:01 Salebot (discuter | contributions) (-25) (bot : révocation de 216.73.72.120 (modification suspecte : -80), retour à la version 65869947 de Klith) (défaite)
- (actu | diff) ☐ ☒ 31 mai 2011 à 20:01 216.73.72.120 (discuter) (+25) (→Chronologie du Moyen Âge) (défaite)
- (actu | diff) ☐ ☒ 30 mai 2011 à 09:44 Klith (discuter | contributions) (+3447) (LiveRC : Révocation des modifications de 83.113.248.116 (retour à la dernière version de 66.110.146.125)) (défaite)

8



## 1.6 Architecture d'extraction de DBpedia

### 1.6.1 Vue d'ensemble

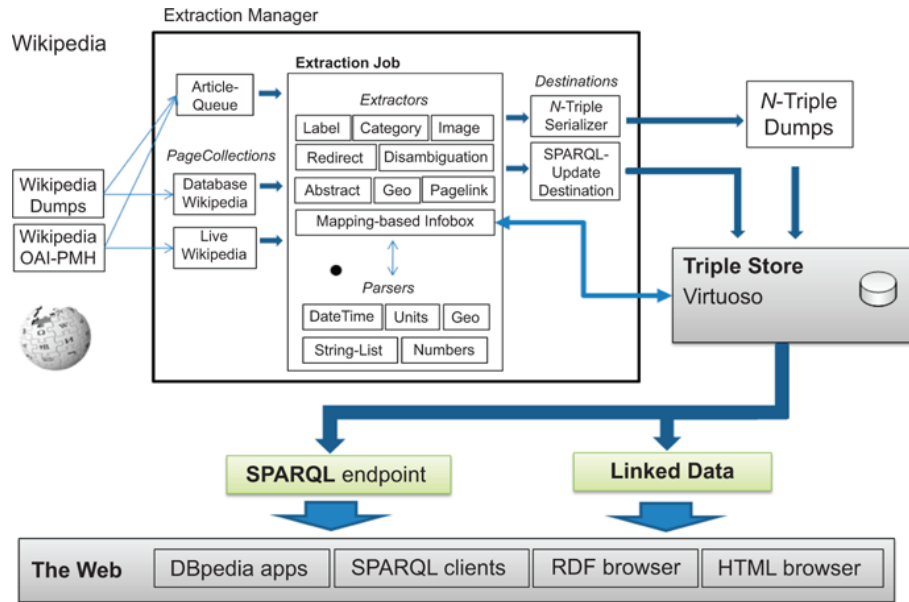


FIGURE 1.8 – Extracteur DBpedia

La figure ci-dessus montre l'architecture du système d'extraction des connaissances dans DBpedia. D'après Morsey et al [?] les principaux éléments du système sont les suivants : *PageCollections* est une abstraction des ressources locales ou distantes des articles de Wikipédia, *Collections* stockent ou sérialisent les triplets RDF extraites, *Extractors* qui transforme un type spécifique de la syntaxe wiki en triplet, *Parsers* soutiennent les *Extractors* en déterminant les types de données, convertit les valeurs entre différentes unités et fractionne les marqueurs dans des listes. L'*Extraction Job* regroupe une collection de pages, extracteurs et destination dans le flux de travail *workflow*. Le noyau de ce système est l'*Extraction Manager* qui gère le processus d'adoption des articles de Wikipédia sur les *Extractors* et donne les résultats à la destination. Le gestionnaire d'extraction *Extraction Manager* gère également la gestion des URI et résout les redirections entre les articles : ce système se compose de 11 extracteurs qui traitent les types des contenus de Wikipédia (*Labels*, *Abstracts*, *Interlanguage links*, *Images*, *Redirects*,

*Disambiguation, External Links, Pagelinks, Homepages, Categories, Geo-coordinates*). Ce framework d'extraction DBpedia est mise en place pour réaliser deux flux : extraction à partir des sources de données (*DataBaseWikipedia page collections*) et une procédure d'extraction directe (*LiveWikipedia page collections with the OAI – PMH protocol*) pour obtenir la version courante des articles.

## 1.6.2 Notre proposition

En analysant les *dumps* DBpedia et en observant l'architecture de cette base de connaissance, nous avons remarqué que pour annoter temporellement les triplets RDF de DBpedia il est plus intéressant d'extraire l'ensemble des propriétés dans DBpedia, puis de trouver des faits qui ont un trait avec le temps, donner une liste de couples à partir de laquelle un expert choisit un couple et valide les résultats de notre algorithme. Par la suite, nous allons présenter en détail notre proposition.

## 1.6.3 Modélisation

Le modèle quaternaire est un modèle qui capte la base du fait avec un indice temporel, l'exemple suivant en montre le principe de ce modèle.

*<politician> elected <president of US> on <date>*

f1, Kennedy elected PresidentOfUSA

f2 :f1, HappenedDate

*<politician> served as <politician office> from <date> to <date>*

f1, Kennedy holdsPoliticalPosition PresidentOfUSA

f2 :f1, startedOnDate

f3 :f1, endedOnDate

*HappenedDate* est utilisée pour dire que le fait est valide que dans ce point du temps.

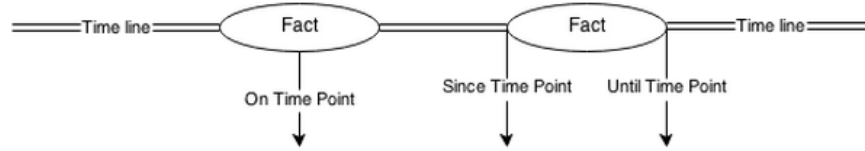


FIGURE 1.9 – Chronologie des événements

Ce modèle est capable d'exprimer la validité temporelle d'un triplet RDF d'une manière à la fois intelligente et lisible par la machine ; on souhaite rattacher au triplet valide que dans un point du temps ou une plage temporelle bien précise une étiquette temporelle adéquate comme le montre la figure suivante.



FIGURE 1.10 – Modélisation quadruplet

On s'intéresse particulièrement au format N-Quads comme format de sortie de notre algorithme. Les quadruplets vont être formalisés de la manière suivante :

$\langle s, p, o, t \rangle$  : un sujet, prédicat, objet avec un point de temps.  
 $\langle s, p, o, [t1, t2] \rangle$  : de même avec une intervalle de temps.

#### 1.6.4 Notre hypothèse

Après une observation approfondie dans les sources de données dans DB-pédia, nous avons repéré des relations logiques entre des propriétés comme (*beatifiedBy*, *beatifiedDate*). Nous avons trouvé plusieurs propriétés qui ont un lien logique entre eux, les relations temporelles ont comme objet un point du temps particulier et partagent le même sujet ou la même ressource avec une autre propriété. Durant cette étude nous avons essayé de valider cette hypothèse :

```
if (x propTemp t) and (x propWithToken z) then
    (x propWithToken z) t
```

*propTemp* est une propriété DBpédia contenant un indice temporel (Year, Date).  
*propWithToken* est une propriété DBpédia avec un motif rattacher.  
*t* est l'annotation temporelle du triplet  $(x, propWithToken, z)$ .

Nous avons présenté cette hypothèse sous forme d'une requête SPARQL. Cette requête interroge l'ensemble des ressources sur DBpédia et retourne des résultats si c'est possible. Notre hypothèse porte principalement sur le fait d'annoter temporellement les ressources de DBpedia en utilisant en essayant de repérer deux triplets portant sur un même sujet et permettant à les relier dont le but est d'avoir un quadruplet valide. La liste des couples (*PropTemp*, *PropWithToken*) est donnée comme sortie d'une procédure d'extraction intelligente de l'ensemble des propriétés de DBpédia.

## 1.7 Notre choix

### 1.7.1 Architecture du système

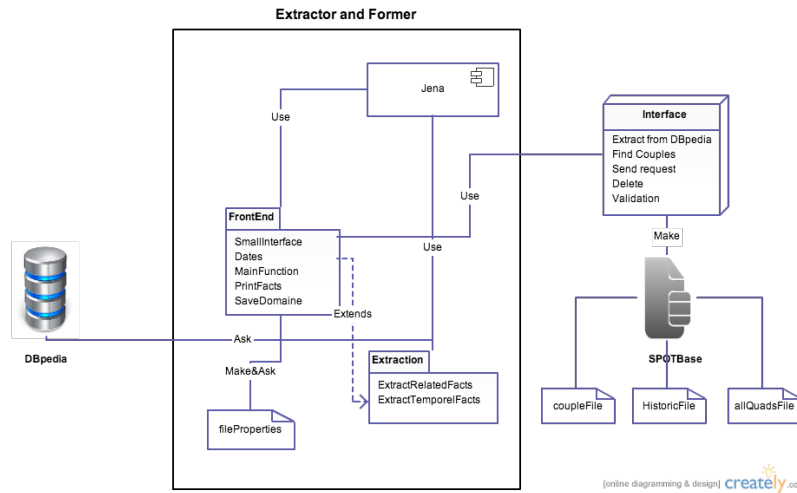


FIGURE 1.11 – Architecture de l'application

L'architecture de notre application se repose principalement sur celle de DBpédia. En premier lieu, nous interrogeons DBpédia pour avoir une liste

de propriétés. En effet, nous pouvons prendre la liste de toutes les propriétés en interrogeant *Virtuoso SPARQL Query Editor*<sup>3</sup> avec la requête

```
select distinct ?P where {?S ?P ?O},
```

mais on se limite aux propriétés de DBpedia qui ont la forme suivante

```
?S rdfs:domain ?O
```

*rdfs : domain* est une instance de *rdfs : Property* qui est utilisé pour indiquer que toute ressource qui possède une propriété donnée est une instance d'une ou plusieurs classes. Le triplet précédant indique que, *S* est une instance de la classe *rdf : Property*, *O* est une instance de la classe *rdfs : Class* et les ressources désignées par les sujets des triplets dont le prédicat est *S* sont des instances de la classe *O*. Lorsque une propriété *S* a plus d'une propriété *rdfs : domain*, les ressources indiquées par les sujets des triplets avec prédicat *S* sont des instances de toutes les classes indiquées par les propriétés *rdfs : domain*. *rdfs : domain* peut être appliqués à lui-même. *rdfs : domain* de *rdfs : domain* est la classe *rdfs : Property*. Cela veut dire que toute ressource avec une propriété *rdfs : domain* est une instance de *rdf : Property*.

Ensuite, nous avons choisi de stocker l'ensemble de propriétés dans un fichier pour ne pas avoir des contraintes de mémoire (stockage dans la mémoire vive) et pour ne pas interroger la base de connaissance à chaque fois. Cette procédure se fait une seule fois l'hors du premier lancement de l'application et elle ne sera plus nécessaire après, car il suffit de spécifier le nom du fichier des propriétés DBpédia que nous avons utilisé l'hors de la première exécution, mais nous avons mis la possibilité d'extraction et mise à jour de ce fichier parce qu'il se trouve que DBpédia change quotidiennement et il y a des propriétés qui s'ajoutent au fur et à mesure à cette base de connaissance. Puis, à partir de ces propriétés, nous avons implémenté un algorithme d'extraction qui à comme sortie une liste de couples de propriétés (PropriétéTemporelle, PropriétéReliée). Dans l'application, nous avons choisi de prendre l'avis d'un expert pour valider les résultats de notre algorithme à travers une liste labellisée d'une partie des quadruplets que nous avons réussi à former et à extraire automatiquement dans un *output Textarea*. Nous avons écrit notre hypothèse de base sous forme d'une requête SPARQL de la manière suivante :

---

3. <http://dbpedia.org/sparql>

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbp:<http://dbpedia.org/ontology/>
SELECT CONCAT(?label1, relatedProp , ?label2, ' : ', ?date)
WHERE {
    ?S    dbp:relatedProp    ?O;
    dbp:tempProp    ?date;
    rdfs:label    ?label1.
    ?O    rdfs:label    ?label2.
    FILTER(lang(?label1)='en' && lang(?label2)='en')}

```

- *tempProp* est une propriété temporelle proposé.
- *relatedProp* est une propriété reliée à *tempProp* partage avec elle un même motif “Token”.

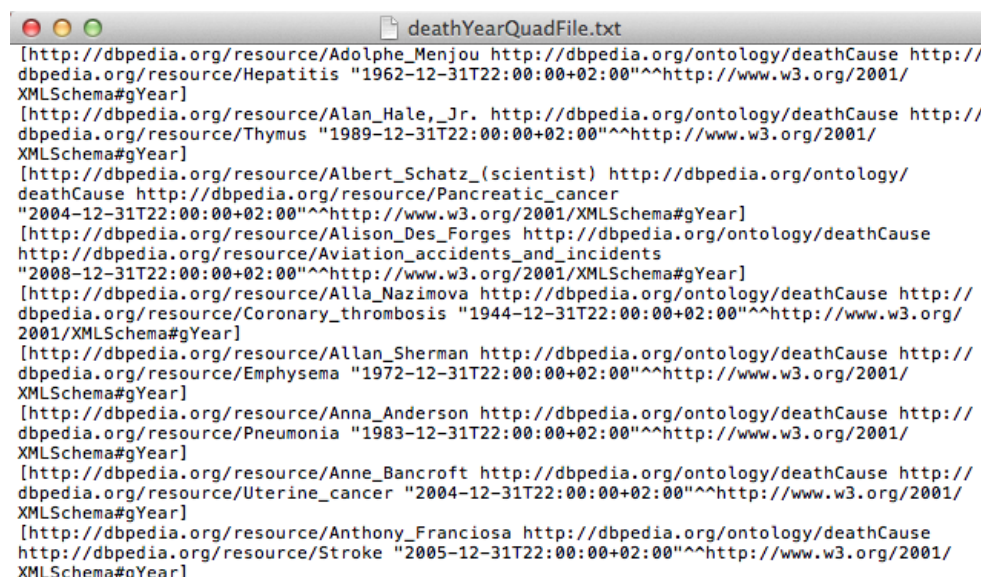
L’objectif de cette procédure est de permettre à l’expert de valider ou ne pas valider la logique de la représentation des quadruplets. Enfin, la validation des résultats permet de stocker l’ensemble des résultats (triplets annotés) dans un fichier portant les labels du couple et un autre fichier “allQuadsFile” contenant tous les quadruplets validés. Par la suite, notre algorithme fait automatiquement l’appel à un fichier CSV d’historique qui a forme suivante (*attribute,tempAttribute,boolean,date\_Exploration,keep,file*) Où *attribute* et *tempAttribute* représentent le couple de propriétés DBpédia, *boolean* peut être 0 ou 1 qui désignent respectivement la volonté de l’expert de valider ou ne pas validé les résultats, *date\_Exploration* est la date de l’exploration. *keep* représente le nombre de quadruplets que nous avons formé à partir du triplet et *file* est le nom de fichier dans lequel nous avons stocké les résultats. Ce fichier nous permet d’avoir une vision globale sur les résultats de notre étude. L’ensemble de fichiers seront stockés dans un dossier forment une base de données de sortie que nous avons appelé SPOTBase.

## 1.7.2 Analyse et discussion

Nous avons réussi à former 305 couples de propriétés, mais nous pouvons encore restreindre ce nombre. Dans notre méthode, nous avons choisi d’extraire même les propriétés reliées aux propriétés temporelles qui contiennent un motif similaire et non pas seulement qui sont identique au suffixe d’une propriété temporelle. Cela a été dans la mesure d’augmenter le nombre de

nos données de test pour avoir une vision globale sur les propriétés DBpédia et analyser par la suite les différents résultats/possibilités.

Avec certains couples, nous avons eu des très bons résultats, par exemple avec le couple (*deathCause, deathYear*) nous avons réussi à formé 2766 quads. La figure ci-dessous montre le format de la sortie de notre algorithme.



```

[http://dbpedia.org/resource/Adolphe_Menjou http://dbpedia.org/ontology/deathCause http://
dbpedia.org/resource/Hepatitis "1962-12-31T22:00:00+02:00"^^http://www.w3.org/2001/
XMLSchema#gYear]
[http://dbpedia.org/resource/Alan_Hale,_Jr. http://dbpedia.org/ontology/deathCause http://
dbpedia.org/resource/Thymus "1989-12-31T22:00:00+02:00"^^http://www.w3.org/2001/
XMLSchema#gYear]
[http://dbpedia.org/resource/Albert_Schatz_(scientist) http://dbpedia.org/ontology/
deathCause http://dbpedia.org/resource/Pancreatic_cancer
"2004-12-31T22:00:00+02:00"^^http://www.w3.org/2001/XMLSchema#gYear]
[http://dbpedia.org/resource/Alison_Des_Forges http://dbpedia.org/ontology/deathCause
http://dbpedia.org/resource/Aviation_accidents_and_incidents
"2008-12-31T22:00:00+02:00"^^http://www.w3.org/2001/XMLSchema#gYear]
[http://dbpedia.org/resource/Alla_Nazimova http://dbpedia.org/ontology/deathCause http://
dbpedia.org/resource/Coronary_thrombosis "1944-12-31T22:00:00+02:00"^^http://www.w3.org/
2001/XMLSchema#gYear]
[http://dbpedia.org/resource/Allan_Sherman http://dbpedia.org/ontology/deathCause http://
dbpedia.org/resource/Emphysema "1972-12-31T22:00:00+02:00"^^http://www.w3.org/2001/
XMLSchema#gYear]
[http://dbpedia.org/resource/Anna_Anderson http://dbpedia.org/ontology/deathCause http://
dbpedia.org/resource/Pneumonia "1983-12-31T22:00:00+02:00"^^http://www.w3.org/2001/
XMLSchema#gYear]
[http://dbpedia.org/resource/Anne_Bancroft http://dbpedia.org/ontology/deathCause http://
dbpedia.org/resource/Uterine_cancer "2004-12-31T22:00:00+02:00"^^http://www.w3.org/2001/
XMLSchema#gYear]
[http://dbpedia.org/resource/Anthony_Franciosa http://dbpedia.org/ontology/deathCause
http://dbpedia.org/resource/Stroke "2005-12-31T22:00:00+02:00"^^http://www.w3.org/2001/
XMLSchema#gYear]

```

FIGURE 1.12 – Fichier de Quadruplet DeathYear et DeathCause

Il se trouve aussi qu'il y a des couples qui valident notre hypothèse mais qui ne donnent pas de résultats. Il se peut que les deux triplets ne partagent pas le même sujet comme {(wineRegion, wineYear),(whaDraft, whaDraftYear),(areaCode, areaDate), etc...}

## 1.8 Outils

### 1.8.1 Application Java

Le choix de développer le logiciel sous forme d’une application *Java* était un choix personnel et qui s’explique de nombreuses manières. Premièrement, la maîtrise de ce langage de programmation me permet d’utiliser des différentes structures de données et d’explorer la documentation de certaines méthodes plus facilement. De plus, il existe plusieurs sources de documentation sur le Web. En outre une large communauté aide à répondre aux questions si jamais on rencontre des difficultés. Il existe une version Java de DBpédia et cela permet plus facilement d’intégrer mon application open source et disponible sur mon compte<sup>4</sup> github. Enfin, nous avons utilisé les librairies *Jena* qui sont aussi écrites en Java.

### 1.8.2 Jena

Jena<sup>5</sup> est un *Framework* open source écrit en Java pour construire des applications dans les domaines du *LinkedData* et le Web sémantique. Nous avons utilisé les différentes librairies de ce *Framework* pour interroger DBpédia avec des requêtes SPARQL. *Jena* est composé de plusieurs programmes différents qui interagissent entre eux pour traiter des données écrites en RDF. *Jena* fournit un support pour le langage de définition d’ontologies (OWL). Ce *Framework* se compose des différentes API RDF, Ontology et SPARQL. Une couche interface d’application et une troisième couche pour le stockage. La figure ci-dessous présente en détaille les différentes composantes de Jena.

---

4. <https://github.com/metanote/Extraction>

5. <https://jena.apache.org/>



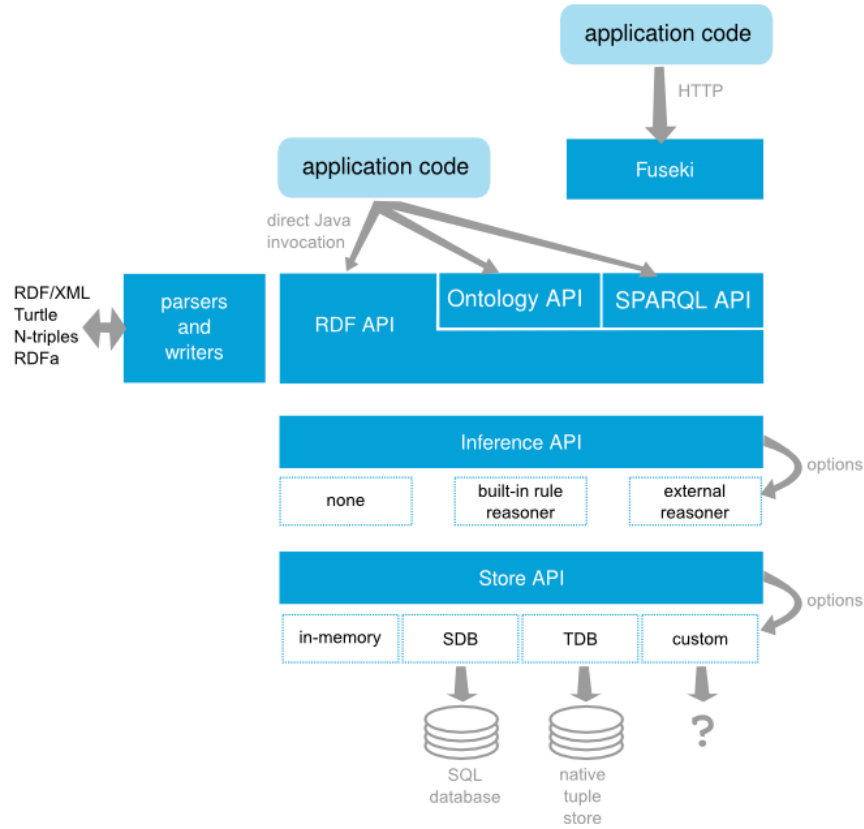


FIGURE 1.13 – Jena Interaction entre les différents API

### 1.8.3 Résultats et Validation

Durant cette étude, nous avons essayé d’annoter temporellement des triplets DBpedia. Nous avons réussi à former automatiquement un nombre important de quadruplet à partir de couple de propriétés qui valide notre hypothèse. Pour certaines propriétés, notre programme prend une quinzaine de minutes des fois pour donner des résultats. Sachant qu’on utilise une machine sous OS x avec un processeur 2 GHz Intel Core i7 et 8 GO de mémoire. Cela est dû au nombre gigantesque de triplets qu’on interroge.

Nous avons stocké plus que 106998 quads que nous avons réussi à former dans le fichier *AllQuadsFile*. Nous avons remarqué que certains couples

valident bien notre hypothèse et donne des excellents résultats. Certes, il y a d'autres couples ne donnent aucun résultat. Le problème, c'est que toutes les propriétés DBpédia ne suivent pas toujours la même logique de représentation. Si tel n'était pas le cas, nous pouvons avoir beaucoup plus de résultats. Dans le Web sémantique, nous avons remarqué qu'il est très important de mettre des conventions pour la représentation des données. Cela permet non seulement d'utiliser les triplets existants, mais aussi de mettre des hypothèses permettant de construire des travaux au dessus de ce qui existait. Le Web sémantique évolue s'il se repose sur une structure de métadonnées générique, claire et réutilisable. Dans cette étude, nous avons traité des triplets DBpédia. Nous avons réussi à implémenter une solution pour annoter des triplets qui ont un trait avec le temps et nous avons mis ces triplets sous la forme quadruplet. Nous avons manipulé une partie de triplets de la base de données d'entrer pour construire notre base de données quadruplets de sortie que nous avons appelé SPOTBase.

## 1.9 Résumé

Dans cette section, nous avons présenté notre approche à travers un système d'annotation temporelle des triplets dans DBpédia que nous avons développé. Dans un premier lieu, nous avons effectué une étude préliminaire en analysant les besoins et en étudiant l'architecture de la base de connaissance. Par la suite, nous avons proposé une solution que nous avons implémenté sous forme d'un prototype fonctionnel.

Dans la prochaine section, nous présenterons les perspectives et les pistes d'amélioration possibles. On conclut ce travail avec un récapitulatif qui résume la richesse de cette étude et son apport considérable dans le domaine de l'ingénierie et la recherche.