N° d'ordre :

REPUBLIQUE ALGERIENNE DEMOCRATIQUE & POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR & DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DJILLALI LIABES - SIDI BEL ABBES
FACULTE DES SCIENCES EXACTES
DEPARTEMENT D'INFORMATIQUE

THESE DE DOCTORAT EN SCIENCE

Présentée et soutenue par
BENYAHIA Kadda

Annotation sémantique des documents

Dirigée par : Pr. LEHIRECHE Ahmed

Soutenue le **13/04/2017** devant le jury :

Mr. ELBERRICHI Zakaria	Professeur à l'université de sidi bel abbès	Président
Mr.BENSLIMANE Sidi Mohamed	Professeur à l'ESI de sidi bel abbès	Examinateur
Mr. HAMOU Mohamed Reda	M.C.A à l'université de saida	Examinateur
Mr. BENSABER Djamel Amar	M.C.A à l'ESI de sidi bel abbès	Examinateur
Mr. KESKES Nabil	M.C.A à l'ESI de sidi bel abbès	Examinateur

Année universitaire : 2016-2017

Dédicaces

A ma grande et petite famille

Remerciements

J'aimerais tout d'abord exprimer ma gratitude et mes sincères appréciations :

A mon directeur de thèse le professeur **Lehireche ahmed** pour sa disponibilité, ses critiques et ses conseils toujours opportuns ainsi que pour sa grande gentillesse et de m'avoir donné toute cette confiance et d'avoir cru à mes compétences.

Aux membres de jury pour leur disponibilité et acceptation d'examiner et de rapporter mon travail.

A toutes les personnes qui m'ont aidée durant mon processus de recherche et d'écriture de cette thèse.

Résumé

Le nombre des documents sur le web s'accroit de jour en jour, et la localisation des documents est devenu une lourde tâche surtout s'il s'agit de la recherche d'un contenu. Ajouter une couche sémantique aux documents c'est l'une des méthodes qui donne aux documents plus de sémantique, et alors la recherche devient un sens pas un terme. Donc un document doit être décrit par une liste de concepts reliés par des relations, c'est l'annotation sémantique.

Dans cette thèse, Nous nous intéressons à une approche d'annotation sémantique des documents pédagogiques sur le web. Cette approche vise à annoter un document par contenu et par contexte ; par contenu, le document sera représenté par des mots clés qui seront ensuite reliés à des concepts d'ontologie, et par contexte, puisqu'un document cite généralement d'autres documents, nous propageons les annotations des références pour annoter le document citant. Nous appliquerons ensuite un module de validation qui consiste à rendre nos annotations cohérentes.

Mots clés : annotation sémantique, métadonnées, recherche d'information, Ontologie, validation

Abstract

The number of documents on the web is growing day by day, and the location of the documents has become a difficult task especially if it comes to looking for content. Add a semantic layer to words of documents is one of the methods giving more semantics to the documents, and then the research becomes a meaningful, not just words. So a document must be described by a list of concepts linked by relations, it is the semantic annotation. In this thesis, we present a semantic annotation approach of pedagogic documents on the web.

This approach aims to annotate a document by content and context, by content we represents documents by keywords that are connected to the ontology's concepts. By context, as documents cite generally other documents, we propagate the annotations of references to annotate the citing document. We then apply a validation module, which consists to make our annotations consistent

Key words: Semantic annotation, metadata, information retrieval, ontology, validation.

ملخص

عدد الوثائق على الأنترنت يتزايد يوما بعد يوم و عملية البحث أصبحت مهمة صعبة خاصة إذا ما تعلق البحث بالمحتوى. إضافة طبقة دلالية لمحتوى الوثائق هي واحدة من طرق إعطاءها أكثر دلالة، ومن ثم يصبح البحث بالمعنى لا بالمصطلحات. لذلك يجب أنْ توصف الوثيقة بقائمة من مفاهيم ترتبط بعلاقات معنوية، وهذا ما يسمى بالشرح الدلالي.

في هذه الأطروحة، نقدم منهج للشرح الدلالي للوثائق على شبكة الإنترنت. ويهدف هذا النهج إلى إضافة شروحات للوثيقة من حيث المحتوى والسياق. من حيث المحتوى إذ نقوم بتمثيل محتوى المستند بمفردات يتم ربطها بعد ذلك إلى مفاهيم الأنطولوجيا. و من حيث السياق،اعتمادا على مفهوم أنْ كل وثيقة علمية تحتوي على عدة وثائق مرجعية أخرى، و بالتالي نقوم بتوريث الوثيقة كل شروحات الوثائق التي ذكرتها كمراجع. في آخر خطوة نقوم بتأكيد صحة كل الشروحات الناتجة عن طريق ازالة التكرارات و التناقضات داخل قاعدة الشروحات

الكلمات المفتاحية :الشرح الدلالي ، الفوقية ، البحث عن المعلومات ، أنطلوجيا ، التحقق من الصحة

Table des Matières

Liste des Figures

Liste des Tableaux

Introduction générale

« Le fléau de mon existence est de faire des choses que
je sais que l'ordinateur pourrait faire pour moi » Dan Connolly,

1. Introduction

Le Web sémantique fournit un moyen d'échange d'information et de savoir en partageant les ressources. En particulier, nous avons cherché à appliquer les technologies du Web sémantique afin d'offrir une plateforme permettant le partage des documents pédagogiques. Les techniques employées dans le Web sémantique afin de partager les ressources nécessitent que celles-ci soient explicitement décrites. Ces données additionnelles de description seront exploitées par les utilisateurs ou par les agents logiciels afin de localiser et d'extraire les documents. Par conséquent, une description sémantique des documents est un concept fondamental du Web sémantique. Cependant, et comme nous allons le voir (percevoir) dans notre cas, le contenu des documents n'est pas toujours bien enrichi pour diverses raisons. C'est pourquoi, nous allons présenter des techniques afin de pallier ce problème en décrivant sémantiquement les documents pédagogiques afin de les rendre exploitable et partageable tout en gardant notre base d'annotation cohérente.

2. Problématique

Un enseignant échange et partage ses ressources et connaissances entre les différents enseignants et apprenants. Il a besoin d'outils pour l'aider à décrire ses documents pour les maintenir à jour quand des évolutions techniques se produisent, et vue au nombre de documents pédagogiques créées sur le Web chaque année qui correspond au nombre de cours et exercices diffusés sur le web qui dépasse largement celui des livres publiés.

Ces documents sont souvent trop longs pour être faciles à lire, en particulier quand les informations importantes sont dispersées dans différentes parties et souvent définis de façon plus ou moins formelle, ils doivent être bien décrits, sinon ils deviennent inexploitables et impossible à retrouver. Donc l'utilisation d'un vocabulaire commun comme ontologies s'avère une solution aux problèmes de partage, et par conséquent l'annotation sémantique des documents en utilisant les ontologies de domaines ajoute une couche sémantique facilitant la tâche de partage d'une part et de la recherche d'autre part.

L'annotation sémantique des documents est la méthode la plus pertinente et la plus prometteuse pour pallier aux problèmes de volatilité et d'hétérogénéité des documents mais elle soulève trois problèmes principaux :

1- Représentation des documents par des mots clés : Un document est décrit par une liste de mots clés qui représente son contenu, ces mots clés plats ne sont reliés par aucune relation, et donc l'utilisation des concepts d'un vocabulaire reliés par des relations rend le document bien décrit et nous permettra de faire une recherche sur son contenu.

2- N'utiliser que le contenu du document génère un manque qui peut être dû aux raisons suivantes :

- Un auteur d'un document peut vouloir faire passer une idée dans un document et de ne pas utiliser les termes de l'ontologie.
- Un document peut contenir beaucoup d'idées et ne pas avoir beaucoup de contenus. Par exemple un document décrivant les présentations d'une conférence va avoir peu de texte, mais de nombreuses références concernant les présentations.
- Le contenu des documents n'est pas toujours disponible mais seulement un ensemble de métadonnées associées.

3- L'annotation dépend des annotateurs et donc peut engendrer une incohérence, puisqu'il peut arriver qu'une ressource soit annotée de différentes manières en utilisant différents vocabulaires et par conséquent les métadonnées ajoutées ne reflètent pas fidèlement le contenu d'un tel document.

3. Contributions

Dans cette thèse, nous présentons une approche d'annotation valide, qui consiste à annoter des documents pédagogiques et de valider la base d'annotation.

3.1 L'annotation par contenu

Un document doit être bien décrit, sinon il peut être inexploitable et difficile à retrouver. L'annotation des documents recouvre ce problème, mais cette tâche reste difficile à effectuer manuellement. Notre contribution consiste à définir une nouvelle approche d'annotation de documents pédagogiques. Un document bien présenté offre aux annotateurs plus de choix, plus de mots clés significatifs à utiliser dans la phase d'annotation. Notre système est basé dans sa phase d'extraction des mots clés des documents sur deux processus, le processus de pondération qui devrait fournir une représentation iconique, compacte et informative du contenu du document, et la mesure de similarité qui calcule le poids d'un mot dans le document. L'utilisation de mots clés plats marque la limite de cette étape et donc l'utilisation d'ontologie dans notre approche offre une certaine efficacité à cette annotation.

3.2 L'annotation par contexte

Quand il est difficile d'accéder au contenu d'un document, la tâche d'annotation devient plus difficile, et par conséquent trouver un mécanisme permettant de le doter par des métadonnées fait l'objet de notre système. Partant du constat qu'un document pédagogique cite généralement autres documents, nous avons proposé de propager les annotations des documents cités sur le document citant et ceci complètement indépendant du contenu.

3.3 La validation

Pour les systèmes d'annotations sémantique, peu de travaux qui traite le problème de validation de ses systèmes. Puisqu'il s'agit de valider des systèmes qui opèrent des raisonnements sur des bases d'annotations qui ont défini d'une manière souvent approximative. A l'heure actuelle, où le web sémantique avec les annotations sémantiques pénètrent tous les domaines, comment admettre que la recette d'une telle annotation ne soit rien d'autre qu'un miraculeux acte de foi qu'un acte de confiance mutuelle entre les annotateurs et les moteurs de recherche qui en prend livraison ?. Quand on sait que les annotations sémantiques seront amenées plus en plus souvent à aider les moteurs de recherche à la prise de décision, la nécessité de penser à des approches de validation devient une lourde tâche.

Nous avons proposé un module de validation qui consiste à rendre la base d'annotation cohérente, ce module se déclenche généralement chaque fois une annotation est créée, il élimine les redondances, et recherche l'incohérence à fin de la corriger.

4. Déroulement de la thèse

Dans cette thèse, nous avons traité principalement le problème d'annotation sémantique, que nous avons complété par une proposition d'une approche de validation pour rendre notre base d'annotation cohérente.

Le plan de cette thèse s'organise autour d'une progression de la réflexion, partant de la problématique telle qu'elle est traitée dans l'état de l'art (cf. partie 1) vers les solutions opérationnelles qui ont été conçues, en passant par la description de notre contribution, les expérimentations et les résultats. (cf. partie 2).

La première partie présente une vue d'ensemble des différents champs de recherche concernés par notre problématique, commençant par un état de l'art sur le web sémantique son architecture, ses différents langages et les différents ressources Terminologiques ou Ontologiques (cf. Chapitre 1).

Une présentation des annotations et annotations sémantiques fait l'objet du deuxième chapitre (cf. Chapitre 2). Nous avons passé à une présentation guidée par la perspective de l'annotation sémantique, à savoir les différents travaux portants sur les annotations sémantiques selon deux axes, par le contenu du document et celui de l'annotation en utilisant le contexte du document (cf. Chapitre 3). Cette première étude nous permet d'exposer les notions essentielles pour la compréhension du domaine de recherche et de mettre l'accent sur les points clefs liées à notre problématique. Puis, nous abordons la partie contribution.

Dans la deuxième partie, nous présentons notre contribution d'annotation sémantique de documents. Dans le chapitre 4, l'approche d'annotations des documents en se basant sur le contenu et sur le contexte est détaillée. Ceci permet d'annoter un document par son contenu en représentant le document par un ensemble des mots clés qui seront reliés à une ontologie de domaine et par contexte sans connaitre son contenu suivi par l'étape de validation pour rendre notre base d'annotation cohérente. Le chapitre qui suit (cf. chapitre 5), est consacré à l'expérimentation de notre approche ainsi que l'évaluation des résultats. Afin de tester notre approche, nous avons implémenté un outil qui regroupe les différentes phases de notre approche. Nous présentons l'expérimentation effectuée pour évaluer l'approche proposée.

Le web sémantique

Les gens qui disent que le web sémantique est une mauvaise vision, ou impossible à mettre en œuvre ont clairement oublié le point : le Web sémantique devrait être un Web qui nous aide à faire des choses qui seraient mieux couverts par les machines. En fait, le Web sémantique est ce que nous voulons qu'il soit. Dans ce chapitre, nous présentons une description de quelques concepts dans le web sémantique.

"The first step is putting data on the Web in a form that machines can naturally understand, or converting it to that form. This creates what I call a Semantic Web—a web of data that can be processed directly or indirectly by machines."

—Tim Berners-Lee, Weaving the Web, Harper San Francisco, 1999

1.1 Introduction

Le développement du World Wide Web est une grande réussite en ce qui concerne le nombre d'utilisateurs et la quantité d'informations offerts par la WWW. Cependant, la plupart des informations disponibles doivent être interprétées par les humains ; Le support machinable est assez limité. Afin de se débarrasser de cette limitation, Tim Berners-Lee, l'inventeur de la WWW, a inventé la vision du Web sémantique : rendre le contenu de la WWW accessible et interprétable par les machines. Dans le présent chapitre nous présentons le web sémantique, son architecture, ses langages et ses différentes ressources.

1.2 Contexte

Le World Wide Web a été inventé par «Sir Tim Berners-Lee » en 1989.La technologie clé du web d'origine du point de vue de l'utilisateur final, de toute façon était l'hyperlien, un utilisateur peut cliquer sur un lien et immédiatement aller au document identifié dans ce lien.

Le grand avantage du Web 1.0 réside dans l'abstraction des couches physiques de stockage et de mise en réseau impliquées dans l'échange d'informations entre deux machines. Cette percée a permis aux documents d'apparaître être directement connectés les uns aux autres. Cliquez sur un lien et vous êtes là, même si ce lien va vers un autre document sur une machine différente sur un autre réseau sur un autre continent !

De la même façon que Web 1.0 a éliminé les couches réseau et physique, le Web sémantique élimine les couches de documents et d'applications impliquées dans l'échange d'informations. Le Web sémantique connecte les faits, de sorte que plutôt de lier à un document ou une application spécifique, vous pouvez se référer à une information spécifique contenue dans ce document ou application. Si ces informations sont toujours mises à jour, vous pouvez automatiquement profiter de la mise à jour. [DO03]

Le mot sémantique lui-même implique le sens ou la compréhension ; la différence fondamentale entre les technologies du Web sémantique et d'autres technologies liées aux données (comme les bases de données relationnelles ou le World Wide Web lui-même) est que le Web sémantique s'intéresse au sens et non à la structure des données. Cette différence fondamentale engendre une vision complètement différente de la manière dont on pourrait aborder le stockage, l'interrogation et l'affichage des informations.

1.3 Le web sémantique

1.3.1 Généralités

Le Web sémantique est une idée de l'inventeur de World Wide Web « Tim Berners-Lee » que le Web dans son ensemble peut être rendu plus intelligent et peut-être même intuitif

sur la façon de servir les besoins d'un utilisateur. Berners-Lee observe que bien que les moteurs de recherche indexent une grande partie du contenu du Web, ils ont peu de capacité de sélectionner les pages qu'un utilisateur veut vraiment ou a besoin [DO03]. Il prévoit un certain nombre de façons dont les développeurs et les auteurs, seuls ou en collaboration, peuvent utiliser des auto-descriptions et d'autres techniques pour que les programmes de compréhension du contexte puissent trouver sélectivement ce que veulent les utilisateurs.

Le Web sémantique est un Web qui comprend des documents ou des parties de documents décrivant des relations explicites entre les choses et contenant des informations sémantiques destinées au traitement automatisé par nos machines. Il fonctionne sur le principe des données partagées.

Il y aura de nombreuses couches sur le Web sémantique (cf. Figure 1.1).

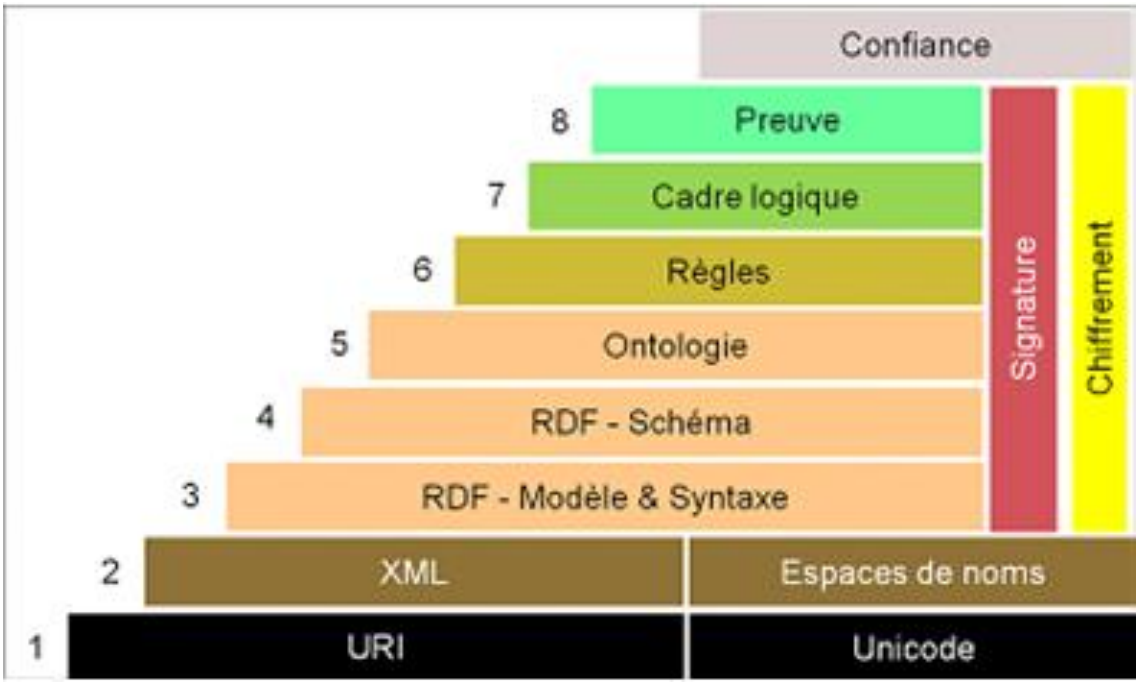

Figure 1.1 Architecture pyramidale du web sémantique [W3C][1]

1.3.2 L'architecture du web sémantique

1.3.2.1 L'URI

L'URI (Uniform Ressource Identifier, identifiant uniforme de ressource) suit les caractéristiques importantes de la WWW existante. C'est une chaîne d'un formulaire standardisé qui permet d'identifier uniquement les ressources (par exemple, les documents).

Un sous-ensemble d'URI est Uniform Resource Locator (URL), qui contient le mécanisme d'accès et un emplacement (réseau) d'un document comme

[1] https://semantiquee.wordpress.com/2014/12/07/architecture-web-semantique

http://www.example.org/. Un autre sous-ensemble d'URI est URN qui permet d'identifier une ressource sans impliquer son emplacement et les moyens de déréférencement. L'utilisation de l'URI est importante pour un système Internet distribué, car il permet une identification compréhensible de toutes les ressources. Une variante internationale de l'URI est l'identificateur de ressource internationalisé (IRI) qui permet l'utilisation de caractères Unicode dans l'identifiant et pour lesquels un mappage à URI est défini.[Dav10]

1.3.2.2 Unicode

Unicode est un standard d'encodage de jeux de caractères internationaux et permet à toutes les langues humaines d'être utilisées (écrites et lues) sur le web à l'aide d'un formulaire standardisé

1.3.2.3 XML

La couche XML (Extensible Markup Language) avec XML namespace et XML schema , assure qu'il existe une syntaxe commune utilisée dans le Web sémantique. XML est un langage de balisage général pour les documents contenant des informations structurées. Un document XML contient des éléments qui peuvent être imbriqués et qui peuvent contenir des attributs et des contenus. Les espaces de nommage XML permettent de spécifier différents vocabulaires de balisage dans un document XML. Le schéma XML sert à exprimer le schéma d'un ensemble particulier de documents XML.[DO03]

1.3.2.4 Les couches RDF M&S (RDF Model & Syntax) et RDF Schéma

RDF (Resource Description Framework), un format de représentation de données de base pour le Web sémantique , est un cadre pour représenter l'information sur les ressources sous forme de graphique. Il était principalement destiné à représenter des métadonnées sur les ressources WWW, telles que le titre, l'auteur et la date de modification d'une page Web, mais il peut être utilisé pour stocker d'autres données. Il est basé sur des triples sujet-prédicat-objet qui forment un graphique de données. Toutes les données du Web sémantique utilisent RDF comme langage de représentation primaire. La syntaxe normative pour la sérialisation de RDF est XML dans la forme RDF / XML. La sémantique formelle de RDF est également définie.[DO03]

RDF lui-même sert de description d'un graphe formé par triples. N'importe qui peut définir le vocabulaire des termes utilisés pour une description plus détaillée. Pour permettre la description normalisée des taxonomies et d'autres constructions ontologiques, un schéma RDF (RDFS) a été créé avec sa sémantique formelle au sein de RDF. RDFS peut être utilisé pour

décrire des taxonomies de classes et de propriétés et les utiliser pour créer des ontologies légères.

1.3.2.5 La couche ontologie

Apporte une évolution car elle assure la description de sources d'information hétérogènes. Ces sources peuvent d'ailleurs formaliser une conceptualisation de choses existantes partagée par plusieurs personnes, voir par toute une communauté. Le rôle de l'ontologie est donc d'aider l'humain et la machine à communiquer, en priorisant sur l'échange de sémantique des informations plutôt que la syntaxe et en utilisant des règles précises. [DO03]

1.3.2.6 La couche Rules (règles)

Offre les moyens de l'intégration, de la dérivation, et de la transformation de données provenant de sources multiples.

1.3.2.7 La couche Logique

Se trouve au-dessus de la couche Ontologie. Certains les considèrent comme étant au même niveau.

1.3.2.8 Les couches Proof (Preuve) et Trust (Confiance)

Permettent de vérifier des déclarations effectuées dans le web sémantique. Si l'on part du web sémantique, qui est le but en soit, on peut voir qu'il dépend des différents agents (logiciels ou machines), qui eux même dépendent d'un service de requête (Query Service dépendant du langage XML), mais aussi de la couche confiance (trust), elle-même dépendant des couches Preuve (Proof) Sécurité (Security) et des règles associées au langage de requête (Rules). Les agents dépendent également de cette couche Règles, qui est basée sur la définition des ontologies, elle-même existant par le biais des seules métadonnées. On retrouve donc ici clairement la démarche stratégique associée au web sémantique.

1.3.3 Architectures et langages

Dans cette partie, nous expliquons de manière détaillée les différents concepts ou piliers du web sémantique

1.3.3.1 Le Langage XML

Le code XML, une recommandation formelle du World Wide Web Consortium (W3C), est similaire au langage HTML (Hypertext Markup Language). XML et HTML contient des symboles de balisage pour décrire le contenu de la page ou du fichier. Le code HTML décrit le contenu de la page Web (principalement les images textuelles et graphiques) uniquement en fonction de la façon dans laquelle doit être affiché ou entrer en interaction.

Dans XML , La structure de données est incorporée aux données; ainsi, lorsque les données arrivent, il n'est pas nécessaire de pré construire la structure pour stocker les données; Il est dynamiquement compris dans le XML. Le format XML peut être utilisé par toute personne ou groupe d'individus ou d'entreprises qui souhaitent partager des informations d'une manière cohérente. XML est en fait un sous-ensemble plus simple et plus facile à utiliser du standard SGML (Standard Generalized Markup Language), qui est la norme pour créer une structure de documents.[DO03]

Le bloc de construction de base d'un document XML est un élément, défini par des balises. Un élément a un début et une étiquette de fin. Tous les éléments d'un document XML sont contenus dans un élément ultra périf connu sous le nom d'élément racine. XML peut également prendre en charge les éléments imbriqués ou les éléments d'éléments. Cette fonctionnalité permet à XML de prendre en charge les structures hiérarchiques. Les noms d'éléments décrivent le contenu de l'élément et la structure décrit la relation entre les éléments.

Un document XML est considéré comme étant «bien formé» (c'est-à-dire capable d'être lu et compris par un analyseur XML) si son format est conforme à la spécification XML, s'il est correctement marqué et si les éléments sont correctement imbriqués. XML permet également de définir des attributs pour les éléments et de décrire les caractéristiques des éléments dans la balise de début d'un élément.

```
<?xml version="1.0"?>
<etudiant>
        <nom>BOURHIS</nom>
        <prenom>Antoine</prenom>
        <phone>0650747719</phone>
        <bureau>
            <numero>AA23</numero>
             <batiment>Batiment A</batiment>
        </bureau>
    <etudiant>
```

Figure 1.2 exemple d'un morceau d'un document XML [BP10]

1.3.3.2 Les métadonnées

Les métadonnées sont des informations structurées qui décrivent, expliquent, localisent ou facilitent la récupération, l'utilisation ou la gestion d'une ressource d'information [DO03]. Les métadonnées sont souvent appelées données sur données ou des informations sur l'information. Le terme de métadonnées est utilisé différemment dans différentes communautés. Certains l'utilisent pour se référer à des informations compréhensibles par

machine, tandis que d'autres l'utilisent uniquement pour les enregistrements qui décrivent des ressources électroniques.[DO03]

D'autres schémas de métadonnées ont été élaborés pour décrire divers types d'objets textuels et non textuels, y compris des livres publiés, des documents électroniques, des outils de recherche d'archives, des objets d'art, des ressources éducatifs et des ensembles de données scientifiques.

Il existe trois principaux types de métadonnées :

-Les métadonnées descriptives : décrivent une ressource à des fins telles que la découverte et l'identification. Il peut inclure des éléments tels que titre, résumé, auteur et mots clés.

- Les métadonnées structurales : indiquent comment les objets composés sont assemblés, par exemple, comment les pages sont commandées pour former des chapitres.

- Les métadonnées administratives fournissent des informations pour aider à gérer une ressource, comme, Quand et comment elle a été créée, le type de fichier et d'autres informations techniques, et qui peut y accéder.

Les métadonnées peuvent décrire les ressources à n'importe quel niveau d'agrégation. Il peut décrire une collection, une ressource unique ou une composante d'une ressource plus importante (par exemple, une Photographie dans un article)

Il y a donc ici une légère nuance entre une métadonnée et une annotation : l'annotation représente une nouvelle donnée attachée à une ressource, alors que la métadonnée est une donnée sur une donnée (Figure 1.3). Un des objectifs dans l'environnement du web Sémantique est de décrire le contenu des ressources en les annotant avec des informations non ambigües, afin de les exploiter par les agents logiciels ou machines.

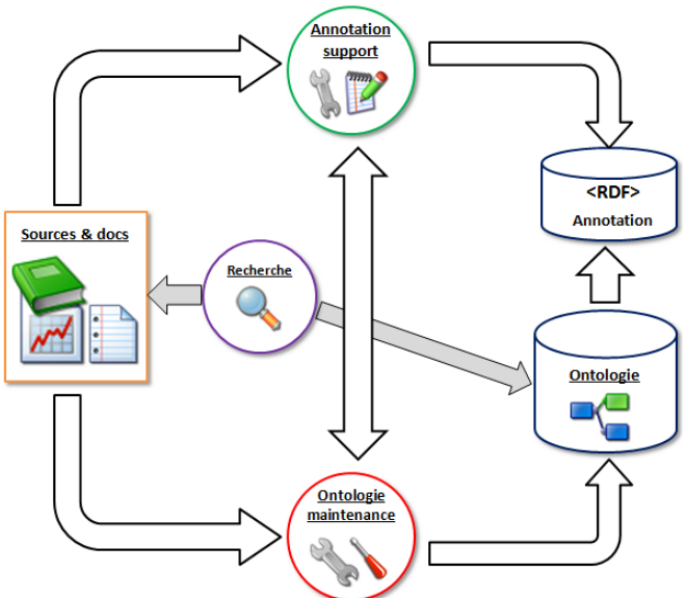

Figure 1.3 Rôle des métadonnées dans un système de gestion de connaissances[BP10]

1.3.3.3 RDF et RDFS

RDF est un modèle standard pour l'échange de données sur le Web. RDF possède des fonctionnalités qui facilitent la fusion de données même si les schémas sous-jacents diffèrent, et il prend en charge l'évolution des schémas au fil du temps sans exiger que tous les consommateurs de données soient modifiés.

RDF étend la structure de liaison du Web pour utiliser les URIs pour nommer la relation entre les choses ainsi que les deux extrémités du lien (cela est généralement appelé un «triple»). En utilisant ce modèle simple, il permet aux données structurées et semi-structurées d'être mélangées, exposées et partagées entre différentes applications.[DO03]

Cette structure de liaison forme un graphe dirigé et marqué, où les bords représentent le lien nommé entre deux ressources, représentées par les nœuds de graphe. Cette vue graphique est le modèle mental le plus facile pour RDF et est souvent utilisée dans des explications visuelles faciles à comprendre.

Au niveau le plus simple, le RDF est un Langage pour décrire les ressources.. Alors que les documents XML attachent des méta-données à des parties d'un Document, une utilisation de RDF consiste à créer des métadonnées sur le document comme entité autonome. En d'autres termes, au lieu de marquer les éléments internes d'un document, RDF capture des métadonnées sur les "externes" d'un document, comme Auteur, la date de création et le type.

Le modèle RDF est souvent appelé «triple» parce qu'il comporte trois parties, ces trois parties sont décrites en termes de parties grammaticales d'une phrase: sujet, prédicat, et l'objet. La figure (figure 1.4) présente les éléments du modèle tripartite et les symboles associés aux éléments lors de la représentation graphique.

Sujet : Dans la grammaire, il s'agit du nom ou de l'expression nominale qui est l'action. Dans la logique, c'est le terme sur lequel quelque chose est affirmé. Dans RDF, c'est le ressource qui est décrite par le prédicat et l'objet.

Prédicat : Dans la grammaire, c'est la partie d'une phrase qui modifie le sujet et comprend l'expression verbale. En d'autres termes, le prédicat nous dit quelque chose sur le sujet. En logique, un prédicat est une fonction d'individus (un type particulier de sujet) à des valeurs de vérité basée sur le nombre d'arguments qu'il a. Dans RDF, un prédicat est une relation entre le sujet et l'objet.

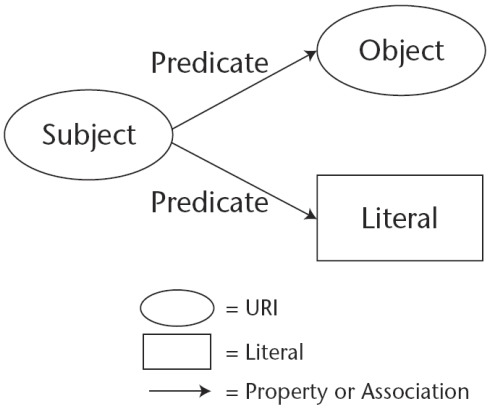

Figure 1.4 Le triplet RDF [DO03]

L'objet : Dans la grammaire, c'est un nom qui est mis en œuvre par le verbe. En logique, un objet est mis en action par le prédicat. Dans RDF, un objet est soit une ressource référencée par le prédicat ou une valeur littérale.

```
<?XML version="1.0"?>
<rdf:RDF  XMLns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
          XMLns:wordnet="http://XMLns.com/wordnet/1.6/"
            XMLns:people="http://www.xulplanet.com/rdf/people/">
<wordnet:Person rdf:about="http://www.xulplanet.com/rdf/people/Karen" people:name="Karen">
  <people:children>
    <rdf:Seq rdf:about="http://www.xulplanet.com/rdf/people/KarensKids">
      <rdf:li>
        <wordnet:Person rdf:about="http://www.xulplanet.com/rdf/people/Sandra" people:name="Sandra"/>
      </rdf:li>
      <rdf:li>
        <wordnet:Person rdf:about="http://www.xulplanet.com/rdf/people/Kevin" people:name="Kevin"/> </rdf:li>
      <rdf:li>
        <wordnet:Person rdf:about="http://www.xulplanet.com/rdf/people/Jack" people:name="Jack"/>
      </rdf:li>
    </rdf:Seq>
  </people:children>
</wordnet:Person>

</rdf:RDF>
```

Figure 1.5 Un extrait RDF /Xml [DO03]

1.3.4 Les Ressources Terminologiques ou Ontologiques (RTO)

Bourigault et al [BA04] ont défini la notion de Ressources Terminologiques ou Ontologiques (RTO) à la croisée des domaines de la Terminologie et de l'Intelligence Artificielle, et plus particulièrement de l'ingénierie des connaissances. Cette notion regroupe plusieurs sortes de ressources, allant des index et glossaires jusqu'aux ontologies en passant par les bases de données lexicales et les thesaurus. Nous allons présenter les trois principales RTO permettant de représenter et de modéliser la connaissance d'un domaine : les taxonomies, les thesaurus et les ontologies.

1.3.4.1 Les Taxonomies

Pour la représentation de l'existant d'une manière formelle, et donc en premier lieu la nécessité de classifier pour étudier et comprendre. L'utilisation des taxonomies est l'un des moyens pour conceptualiser les objets et les classifier hiérarchiquement.[Cha02]

Une taxonomie est définie comme : La classification des entités d'information sous la forme d'une hiérarchie, selon les relations présumées des entités réelles qu'elles représentent. Une taxonomie est une hiérarchie sémantique dans laquelle les entités d'information sont liées soit par la sous-classification de la relation, soit par la sous-classe de la relation.
L'utilisation la plus courante des taxonomies (en fait, la raison principale pour Taxonomies plutôt que d'autres structures de connaissances plus compliquées) est donc naviguer pour obtenir de l'information, surtout lorsque vous avez une idée générale de ce que vous recherchez.

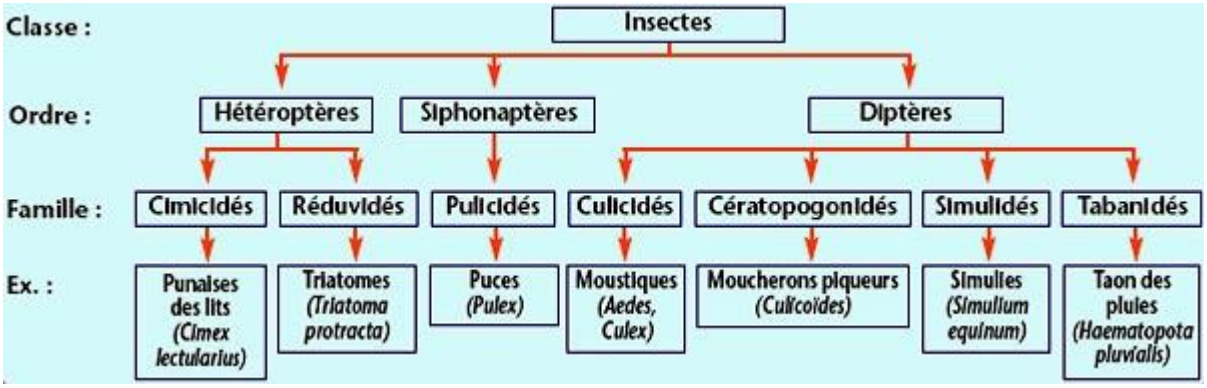

Figure1.6 Extrait d'une taxonomie sur la représentation simplifiée des insectes[2]

Une taxonomie est généralement représentée avec la racine de la taxonomie en haut, comme dans Figure 1.6. Chaque nœud de la taxonomie, y compris la racine, est une information entité qui représente une entité du monde réel. Chaque lien entre nœuds représente une relation spéciale appelée la sous-classification de la relation (si la flèche du lien est Pointant vers le nœud parent) ou est une super-classification de (si le lien La flèche pointe vers le bas sur le nœud enfant). Parfois, cette relation spéciale est définie plus strictement par « *est subclasse de* » ou « *est superclasse de* », où elle est comprise pour signifier que les entités d'information sont des classes d'objets.

[2] http://www.jim.fr/e-docs/00/01/B3/F8

1.3.4.2 Les Thesaurus

Charlet et al [Cha02] a défini un thesaurus comme « un ensemble de termes normalisés fondé sur une structuration hiérarchisée. Les termes y sont organisés de manière conceptuelle et reliés entre eux par des relations sémantiques. Organisé alphabétiquement, il forme un répertoire alphabétique de termes normalisés pour l'analyse de contenu, le classement et donc l'indexation de documents d'information».

Bourigault & al [BA04] définissent un thesaurus comme « un langage documentaire fondé sur une structuration hiérarchisée », sachant qu'un langage documentaire est un « ensemble organisé de termes normalisés, utilisé pour représenter le contenu des documents à des fins de mémorisation pour une recherche ultérieure ». Un thesaurus est donc considéré comme un vocabulaire contrôlé et structuré dans lequel les relations entre les termes du domaine considéré sont clairement spécifiées formant ainsi un réseau terminologique. La structure hiérarchique correspond comme dans les taxonomies à une relation d'hyperonymie sauf qu'elle structure des termes vocabulaire.

Dans la figure 1.7, le terme « Véhicule » a un sens plus général que « voiture »

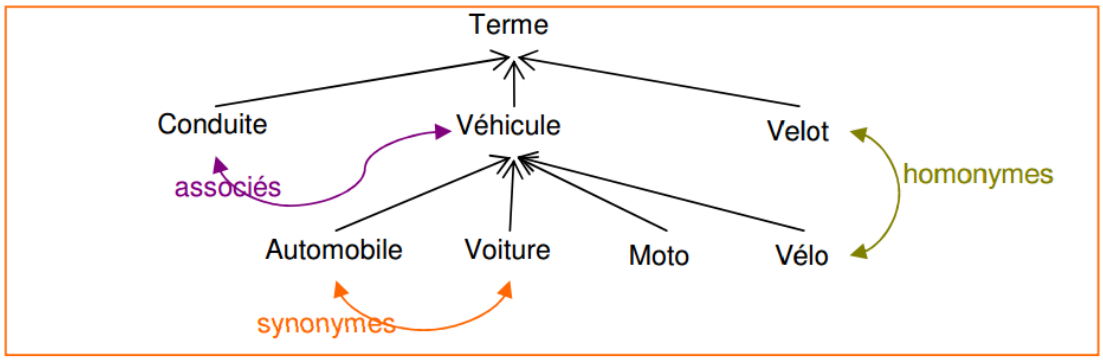

Figure 1.7 Les différentes relations qui composent un thesaurus [Flo07]

- Entre « Voiture » et « Automobile », il existe une relation de Synonymie.
- Entre « Velot » et « Vélo », il existe une relation d'homonymie.

Les thesaurus ne sont pas des ontologies, ils ne fournissent pas de représentation de la connaissance, ils aident dans le cas d'utilisation d'une ontologie, comme pour le cas d'annotation [Her05].

1.3.4.3 Les ontologies

a. Définition

Le dictionnaire nous a donné les définitions suivantes :

1-Une branche de la métaphysique concernée par la nature et les relations de l'être

2-Une théorie particulière sur la nature de l'être ou sur les genres d'existants

Cette définition indique que le terme provient de la philosophie, Partie de la métaphysique qui est l'étude systématique des principes sous-sujet, le plus souvent la nature de l'être et la nature de l'expérience [Cha02]. Souvent la distinction est faite entre «grand O» Ontologie et "Petite o" ontologie. Grand « O » L'Ontologie est la discipline philosophique. Petit « o », est la discipline de l'ingénierie des technologies de l'information qui émergée au cours des dernières années.

Une des premières définitions de l'ontologie a été énoncée par Gruber [Gru93] comme la « spécification explicite d'une conceptualisation ».

Guarino[Gua98] a donné une autre définition : « un vocabulaire spécifique utilisé pour décrire une partie de la réalité, plus un ensemble d'hypothèses explicites concernant la signification prévue de ce vocabulaire».

R. Studer et al. [SB98] ont défini une ontologie comme « spécification formelle et explicite d'une conceptualisation partagée »

Ces définitions comportent d'autres termes :

•Formelle : l'ontologie doit être lisible par une machine, ce qui exclut le langage naturel.

•Explicite : la définition explicite des concepts utilisés et des contraintes de leur utilisation.

• Conceptualisation : rend compte du sens des termes, c'est la définition du modèle abstrait d'un phénomène du monde réel par identification des concepts clefs de ce phénomène.

L'ontologie nous fournit les outils d'organiser les concepts d'un domaine hiérarchiquement en décrivant leurs propriétés sémantiques dans un langage de représentation des connaissances formelles afin de favoriser le partage d'une vue consensuelle sur ce domaine entre les applications informatiques qui en font usage [BA04].

La définition des concepts et la création des relations sémantiques entre eux correspondent au premier niveau à une ontologie. Une définition formelle d'une ontologie est présentée par Handschuh [Han05] (figure 1.8).

Les concepts (« classes »)

Les objets du monde réel quel que soit leurs nature (abstraits / concrets, réels / fictifs, élémentaires / composites) sont représentés par des concepts qui sont organisés en taxonomie par la relation de subsumption. Dans la Figure (figure 1.9), le concept « Film » est une sous-classe de « œuvre artistique ».

A core ontology *is a structure*

$$\mathcal{O} := (C, \leq_C, R, \leq_R, A)$$

consisting of

- *three disjoint sets* C, R *and* A *whose elements are called* concepts , relations *and* attributes *respectively,*
- *a partial order* $\leq_C$ *on* C, *called* concept hierarchy *or* taxonomy,
- *a partial order* $\leq_R$ *on* R, *called* relation hierarchy.

We furthermore have two functions, domain*:* $R \cup A \rightarrow C$ *and* range*:* $R \rightarrow C$.

Figure 1.8 Définition formelle d'une ontologie donnée par Handschuh [Han05]

Les relations

Les interactions entre concepts qui permettent une représentation complexe de la connaissance du domaine sont représentées par des liens sémantiques binaires appelés relations [CB04]. Dans le figure (figure 1.9) ,une relation sémantique « réalise » existe entre les concepts « Personnalité » et « Film ».

Les attributs

Ce sont des caractéristiques particulières permettent de définir un concept de manière unique dans le domaine [CB04]. Leurs valeurs sont littérales, comme une chaîne de caractères ou un nombre entier. Par exemple, dans la figure (figure 1.9), le concept « Film » peut avoir comme attributs : un « titre », une « date de production », etc.

Les instances

Ou les individus font partie de la base de connaissance [Han05]. Ils permettent de stocker les instances des concepts et de relations et les valeurs des propriétés selon le domaine de l'ontologie. Dans la figure (figure 1.9), « le parrain » est une instance du concept (Film).

Les ontologies représentent une composante importante du web sémantique puisque son objectif est la modélisation des ressources web pour les rendre accessibles par tous. Les ontologies donnent des représentations conceptuelles du monde réel, qui un point important dans la démarche proposée par Berner Lee. L'ontologie en tant que résultat d'une conceptualisation d'un domaine, offre une solution idéale pour relier les réseaux sémantiques associés à des ressources web et donc faciliter l'accès à des informations et des applications.

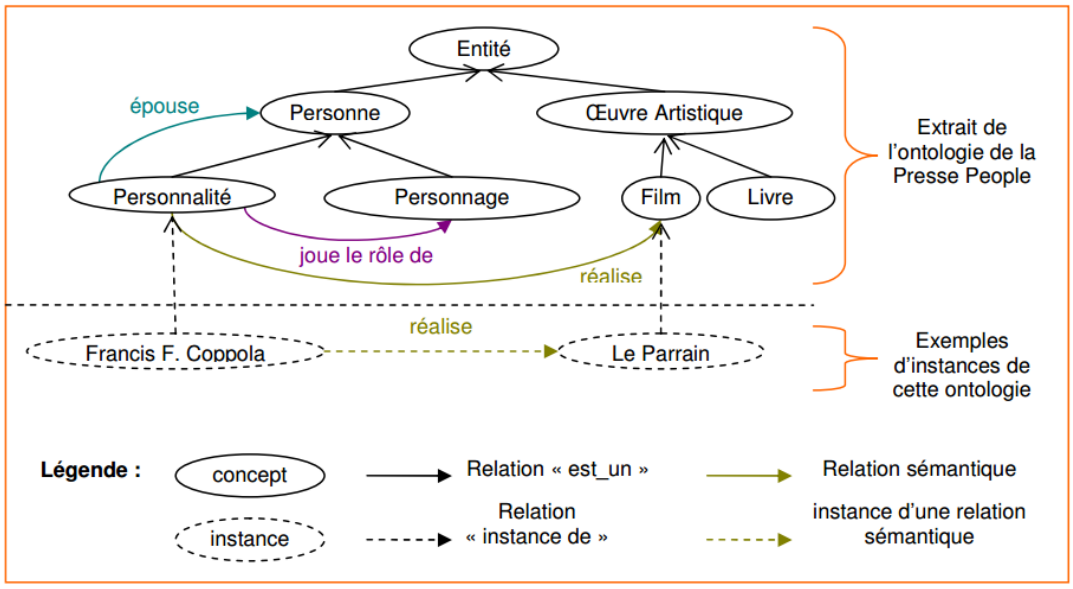

Figure 1.9 Exemple d'une ontologie dans le domaine de la presse « People »[Flo07]

b- Les langages de définition d'ontologies

Les ontologies sont définies par des langages standards :

OWL : (Ontology Web Language) est un langage du Web sémantique conçu pour représenter des connaissances riches et complexes sur les choses, les groupes de choses et les relations entre les choses. OWL est un langage basé sur la logique computationnelle de tel sorte que les connaissances exprimées dans OWL peuvent être exploitées par des programmes informatiques. Les documents OWL, connus sous le nom d'ontologies, peuvent être publiés dans le World Wide Web et peuvent faire référence à d'autres ontologies OWL.

OWL fait partie de la pile de la technologie Web sémantique du W3C, qui comprend RDF, RDFS, SPARQL, ….etc.[DO03]

Le langage OWL dédié aux définitions de classes et de types de propriétés, il est associé à la définition d'une ontologie. Il offre aux machines la possibilité d'interpréter le contenu des ressources web par son vocabulaire et sémantique formelle. Il s'inspire des logiques de descriptions par l'utilisation des constructeurs.

Le langage OWL se compose de trois sous-langages qui symbolisent une expressivité croissante - OWL-Lite,- OWL-DL, OWL-Full.

Les ontologies OWL sont des fichiers texte dont l'extension peut être owl ou rdf puisque le langage OWL repose sur RDF et RDFS, en y ajoutant notamment des nouvelles balises pour gagner en précision.

1.4 Conclusion

Le Web sémantique est un maillage d'information relié de manière à être facilement transformable par les machines, à l'échelle mondiale. C'est un moyen efficace pour représenter des données sur le World Wide Web, ou comme une base de données liée globalement.

Les technologies du Web sémantique dans leur ensemble ont fait des progrès énormes au cours des dernières années :

-Le mouvement Open Linked Data a augmenté massivement chaque année et contient beaucoup plus d'informations que n'importe quelle ressource unique sur le Web.

-Des organisations massives s'appuient maintenant sur les technologies du Web sémantique pour exécuter des opérations quotidiennes critiques.

-Les standards du Web sémantique RDF, SPARQL, OWL et autres - n'étaient que des projets en 2001, mais ils ont été formalisés et ratifiés.

La tâche d'annotation pour le Web Sémantique consiste à enrichir une ressource documentaire par des informations basées sur des représentations de la connaissance plus ou moins formelles.

Dans le prochain chapitre, nous présentons les annotations et les annotations sémantiques.

Les annotations

L'annotation est une phase qui rend les documents exploitables, elle décrit leurs contenus d'une manière à être plus compréhensible par les moteurs de recherche. Ce processus est développé dans ce chapitre.

2.1 Introduction

Avec le développement qui a touché le web et ses technologies, le développement des méthodes et des outils interactifs pour comprendre, manipuler et partager des documents, et mettre en place des services pertinents et performants devient une nécessité, surtout avec le web actuel qui concerne des millions de personnes qui ne se connaissant pas et ayant des différents centres d'intérêt, habitudes et cultures, et que l'information sur le Web est distribuée, volumineuse, évolutive, très "bruitée", et très hétérogène.

L'annotation sémantique semble actuellement l'approche la plus prometteuse pour partager et exploiter l'information sur le Web, elle permette d'associer des notes de lectures aux documents et de partager de l'information. Grâce aux outils d'annotation, le lecteur devient aussi rédacteur.

Dans ce chapitre nous présentons les annotations d'une manière générale et plus précisément les annotations sémantiques.

2.2 Définitions autour des annotations

-Le Petit Robert définit le terme « annotation » comme une « note critique ou explicative qui accompagne un texte – une note de lecture qu'on inscrit sur un livre ».

-Le Dictionnaire de l'Académie Française (9ème édition), définit « annotation » comme un terme dérivé du terme latin annotare, signifiant « noter ; annoter ».

-Le Verbe « annoter » est défini comme « accompagner un texte de notes, de remarques, de commentaires ».

-Le dictionnaire Oxford définit une annotation comme "une note en guise d'explication ou un commentaire ajouté à un texte ou un diagramme".

Il a des usages spéciaux dans différents contextes :

-Dans la programmation du logiciel, une annotation est représentée sous la forme d'un texte de commentaire incorporé dans les codes pour expliquer le programme.

-Dans le dessin mécanique, une annotation est un extrait de texte ou un symbole avec des significations spécifiques qui illustre la partie annotée correspondante.

-Dans la publicité commerciale, une annotation est généralement utilisée comme une sorte de note de bas de détail de certaines restrictions commerciales.

Desmontil[DJ02] a défini une annotation comme une information graphique ou textuelle attachée à un document(Figure 2.1) et le plus souvent placée dans ce document. Cette place est donnée par une ancre. Elle peut prendre plusieurs formes comme :

– des icônes (par exemple pour décrire des avis en utilisant des étoiles, des points d'interrogation...),

– des symboles de liens (pour décrire des associations, des relations entre mots, paragraphes ou chapitres),

– des notes textuelles en marge, en bas de page ou en fin de document repérées dans le texte 2 par des icônes (numéros, étoiles...),

– des mises en forme typographiques (surlignage, soulignage, italique...),

– des redécoupages de texte (à l'aide d'accolades, de numérotation de passages...),

– des images,

– des sons...

Quand nous annotons, nous sommes en train d'expliquer, de commenter un sujet (figure 2.1).Une annotation seule ne fait pas sens, elle est toujours associée à l'objet qui a été annoté. Selon Handschuh [Han05], une annotation constitue un cas particulier d'une métadonnée, elle représente une nouvelle donnée attachée à une ressource. Prié & Garlatti [PG04] définit une annotation comme « un commentaire libre situé à l'intérieur de la ressource documentaire ».

2.3 Types d'annotations

L'annotation d'un document, qu'on appelle aussi indexation [Sal69], [Rij79] consiste à repérer dans son contenu certains mots qui ont une signification dans un contexte donné, et de construire un lien entre les termes sélectionnés et le texte d'origine. Il existe trois types d'indexation :

2.3.1 L'indexation manuelle

Elle assure une correspondance entre les ressources documentaires et l'ensemble des termes à utiliser dans l'indexation. Un très long travail manuel à réaliser par l'indexeur est l'inconvénient principal de ce type d'annotation en plus aux autres démontrés par Salton [Sal86] comme la possibilité que deux indexeurs indexent deux documents identiques avec des termes différents.

2.3.2 L'indexation automatique

Le processus d'indexation est fait par la machine, il est totalement informatisé et réalisé en plusieurs étapes : (i) l'extraction des mots clés qui représente le contenu du document, (ii) le nettoyage de l'ensemble des mots clés sélectionnés et (iii) la pondération des mots, pour affecter un poids aux mots clés.

2.3.3 L'indexation semi-automatique

La phase de sélection de termes est réalisée d'une manière automatique mais la création des liens reste au spécialiste c'est lui qui fait le choix des termes à utiliser dans l'indexation. [EM92]. Le résultat du processus d'indexation classique est une liste de mots indépendants sans aucune relation spécifiée entre eux.

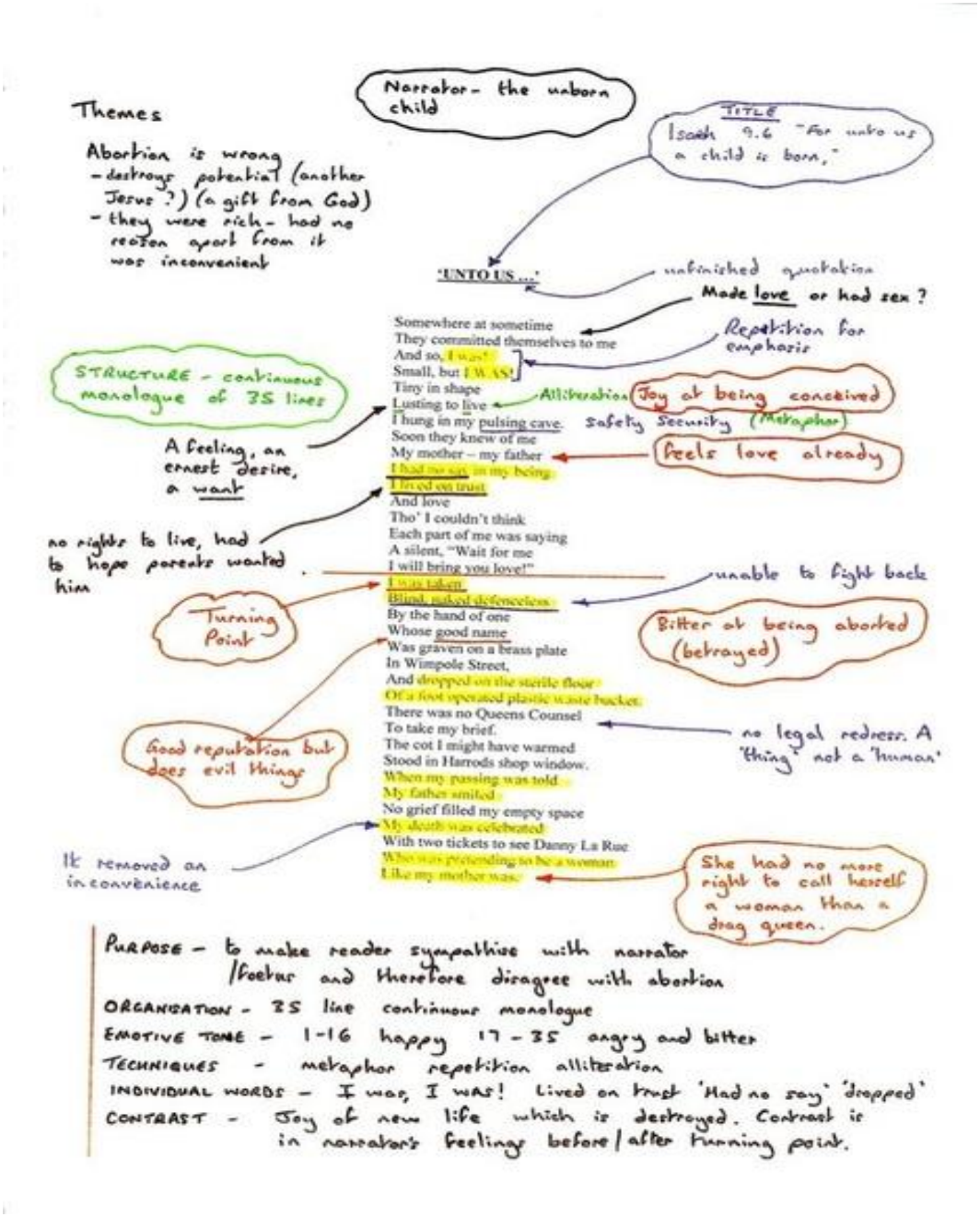

Figure 2.1 : Exemple d'annotation[3]

[3] http://trullsenglish.weebly.com/annotating-text.html

2.4 L'annotation sémantique

Afin de distinguer l'annotation sémantique des autres annotations, plusieurs sortes de classifications sont proposées. Bechhofer et al. [BC02] classent les annotations en trois types: l'annotation textuelle, qui ajoute des notes et des commentaires à un objet; l'annotation de lien, qui étend le type précédent de l'annotation en reliant l'objet à un contenu d'annotation; l'annotation sémantique, qui contient les informations lisibles par l'homme, ainsi que lisibles par machine.

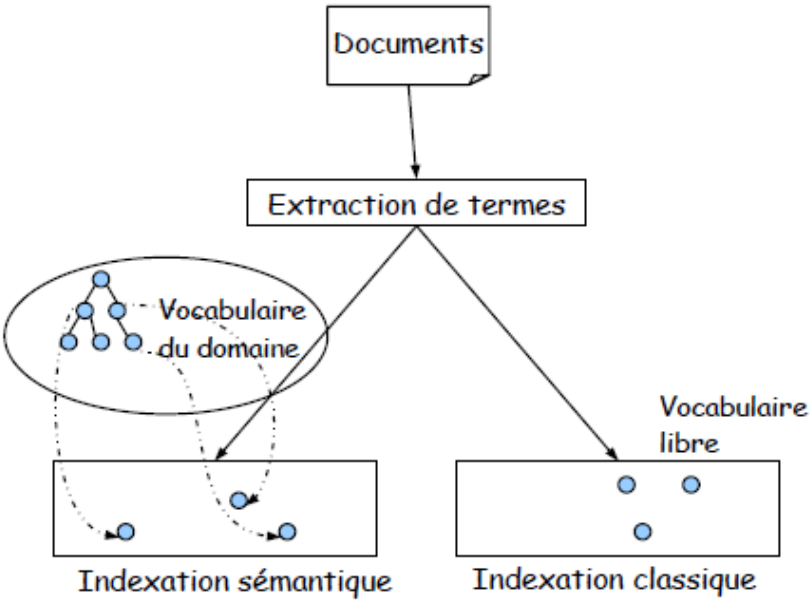

Figure 2.2 : comparaison entre l'indexation classique et l'indexation sémantique [Abr06]

De même, Oren et al. [OM06] ont proposé une autre classification des annotations : l'annotation informelle, qui est exprimée dans un langage informel et ne sont pas lisibles par machine ; l'annotation formelle, qui est lisible par une machine, mais sans conditions ontologiques ; l'annotation ontologique, qui se compose uniquement des termes ontologiques qui sont généralement acceptés et compris dans un domaine spécifique.

Ces classifications identifient deux caractéristiques importantes d'une annotation sémantique : (1) elle est à la fois lisible par l'homme et lisible par machine, et (2) elle contient un ensemble de termes formels et partagés qui peuvent exister pour une communauté humaine et / ou à la machine.

Considérant l'essentiel d'une ontologie [Lin08], qui est un accord commun d'une conceptualisation des termes dans un domaine spécifique, différents chercheurs ont suggéré de nombreuses définitions de l'annotation sémantique liée à une ontologie. Par exemple, Talantikite et al. [TA09] décrit comme «une annotation sémantique est un référent à une

ontologie". Lin [Lin08] considère comme «une approche pour relier les ontologies aux sources d'information d'origine".

Kiryakov et al. [KP04] définissent l'annotation sémantique comme «une génération de métadonnées spécifiques et des schémas d'utilisation qui visent à permettre de nouvelles méthodes d'accès à l'information et d'étendre celles qui existent déjà». À notre connaissance, une annotation sémantique peut être considérée comme un moyen pour effectuer l'enrichissement sémantique de «quelque chose» en utilisant un ensemble de termes bien formalisés et d'un commun accord d'un domaine spécifique, comme ontologies.

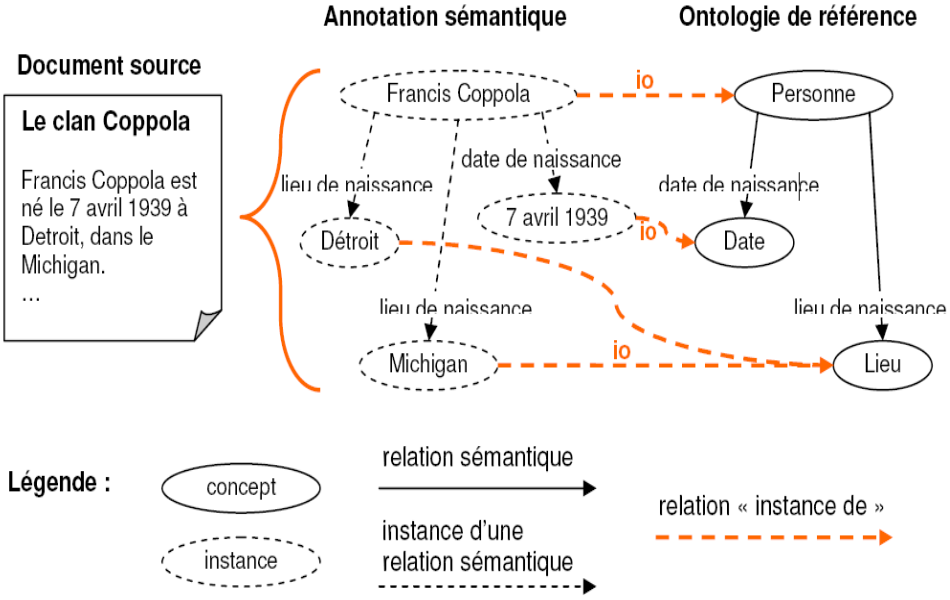

Figure 2.3 : exemple d'une annotation sémantique à l'aide d'une ontologie de référence [Flo07]

Basé sur les supports des ontologies, les annotations sémantiques pourraient être largement utilisées dans de nombreux contextes. Uren et al. [UC06] ont examiné et classé les systèmes d'annotation sémantique existants en quatre catégories: annotation manuelle (les annotations sont créées manuellement par les utilisateurs), annotation automatique (l'annotation est créée avec l'assistant d'automatisation Composants), des environnements d'annotation intégrés (des outils standards tels que Microsoft Word, qui est intégré avec un processus d'annotation) et l'annotation à la demande (Outils qui produisent un service de type annotation, comme la mise en surbrillance de texte).

2.5 Le stockage des annotations et de leurs ressources

Les annotations peuvent être soit « embarquées » soit « débarquées » vis-à-vis de la ressource documentaire source [Hab05].

2.5.1 Les annotations embarquées

Lorsque l'annotation est ajoutée au contenu du document, elle est dite « embarquée». Les systèmes d'annotation, ajoute directement les métadonnées créées à l'intérieur du document source.

2.5.2 Les annotations débarquées

Elle est dite « débarquée » lorsqu'elle est stockée à l'extérieur du document source. Même les liens avec le document à annoter doit être stocké afin de retrouver ensuite toutes les annotations qui concernent une ressource. Les annotations sont généralement stockées sur des serveurs d'annotations (les bases d'annotations) qui peuvent être interrogés afin de retrouver les annotations d'une ressource donnée.

2.6 Conclusion

Les annotations permettent de décrire des documents pour les partager et les rendre mieux exploitable. Les annotations sémantiques, un cas particulier d'annotation qui consiste à représenter les documents par un ensemble des concepts en relations en utilisant un vocabulaire commun qui est généralement une ontologie. Une étude sur les différentes recherches sur les annotations sémantiques sera présentée dans le chapitre qui suit.

Etat de l'art autour des annotations sémantiques

L'annotation sémantique des documents est un processus qui décrit leurs contenus d'une manière à être plus compréhensible par les moteurs de recherche. Une étude sur les travaux de recherche sur les annotations sémantiques est développée dans ce chapitre.

3.1 Introduction

De nos jours, la nécessité de partage d'information et d'interopérabilité des systèmes est devenue de plus en plus omniprésent. De nombreux travaux de recherche ont été menés dans les domaines de l'échange et l'exploitation d'informations. Pour aider à comprendre, manipuler et partager des documents, l'annotation sémantique a gagné de nombreuses attentions et largement utilisée dans différents domaines.

L'enrichissement sémantique des textes est principalement conçu pour aider une machine à «comprendre» la signification des textes annotés et de soutenir des processus automatisés, tels que la navigation de l'information. Bien entendu, pas limité à cela, un grand nombre de recherches ont été proposé.

Nous recueillons des différentes littératures sur les annotations sémantiques des textes, et les classer en fonction de la manière d'annoter par contenu ou par contexte.

Nous avons sondé 52 travaux de recherche d'annotation sémantique des documents, des vidéos, des images et même des web services (y compris les articles de revues, documents de conférence, thèses et rapports PhD), dont nous allons illustrer et discuter ceux annotent des documents textuels dans le présent chapitre. Pour chaque type (par contenu /contexte), nous allons présenter l'analyse de quelques travaux de recherche comme exemples.

Nous présenterons succinctement les deux types d'annotations des textes que nous avons retenus : l'annotation par le contenu, en utilisant uniquement le contenu du document et l'annotation par le contexte qui consiste à utiliser les relations entre les documents

Une étude détaillée de la littérature de l'annotation sémantique des documents est présentée suivie par une comparaison pour identifier les inconvénients existants et de souligner les axes de recherche possibles.

3.2 Annotation des documents par contenu

L'annotation par contenu, utilise uniquement le contenu du document pour le décrire. Il existe plusieurs travaux traitant de l'utilisation de l'annotation sémantique par contenu ; Nous en avons retenus neuf que nous avons estimé pertinents afin de montrer l'utilité d'une telle approche :

- Indexation semi-automatique de documents
- Indexation des documents dans un référentiel métier avec approche ontologique
- Indexation dans un système de recherche documentaire
- Indexation conceptuelle guidée par ontologie pour la recherche d'information.
- Indexation pour la désambigüisation des termes

- Indexation par ontologie de domaine
- MnM ,un outil basé sur l'ontologie
- La plate-forme KIM
- Un framework pour l'annotation sémantique

3.2.1 Annotation semi-automatique de documents [DJ02]

Proposé par Desmontils [DJ02] et [DJM02], un processus d'annotation semi-automatique qui s'appuie sur des techniques issues du traitement automatique des langues et de l'ingénierie des connaissances. La figure 3.1 illustre les différentes étapes de ce processus :

- Pour chaque document ; un index de termes est construit, un poids est associé à chaque terme.
- La construction d'une liste de concepts candidats en utilisant le thésaurus *Wordnet*
- Chaque concept candidat dans le document à un niveau de représentativité qui sera déterminé.
- Un filtre est utilisé en basant sur l'ontologie et la représentativité des concepts. Les documents sont associés aux concepts de l'ontologie.

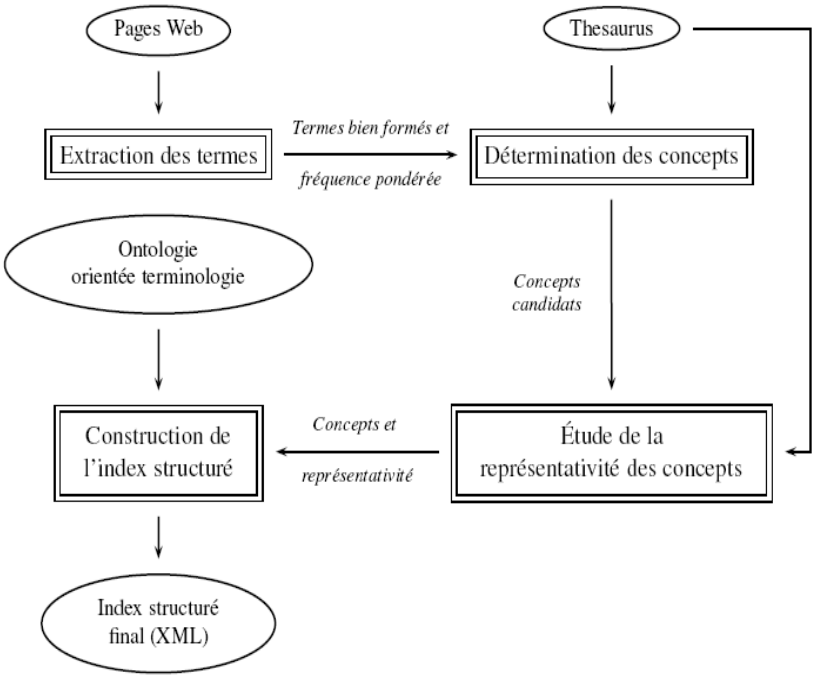

Figure 3.1 Processus d'indexation de Desmontils [DJ02]

3.2.2 Annotation des documents dans un référentiel métier avec approche ontologique [NF04]

Cette approche est motivée par la contrainte que l'indexation d'un document dépend des activités de l'entreprise et non pas des mots clés du document. Elle combine une analyse linguistique du document et une analyse statistique ainsi qu'un traitement sémantique [NF04], [NFF04]. Ses principales étapes sont :

1. Le traitement linguistique qui consiste à extraire les termes composant des documents, en utilisant un outil d'extraction terminologique existant.
2. En se basant sur le référentiel métier (l'ontologie du domaine), le document sera représenté par un ensemble de termes simples et importants.
3. La détermination de l'importance d'un terme dans un document.
4. L'affectation du document au référentiel métier.

3.2.3 Annotation dans un système de recherche documentaire [GM99]

Guarino et al [GM99] ont proposé un mécanisme d'indexation dans un système de recherche documentaire OntoSeek qui utilise l'ontologie Sensus : 50000 nœuds concepts.

L'idée de base est la représentation de contenu du document et des requêtes par un formalisme de graphes conceptuels.

OntoSeek a montré l'efficacité d'indexation par ontologie qui améliore la phase de recherche.

3.2.4 Annotation conceptuelle guidée par ontologie pour la recherche d'information [Baz05]

Dans Ce modèle, Baziz[Baz05] représente le contenu sémantique des documents par son projection sur une ontologie linguistique générale.

Les étapes de cette approche sont les suivantes :

1. L'extraction des termes du document pouvant représenter les concepts de l'ontologie.
2. Un calcul de similarité entre les concepts sera appliqué en utilisant différentes relations (synonymie, hyperonymie,...).
3. La construction du noyau sémantique qui représente mieux le document correspond au score maximum calculé pour ce concept.

3.2.5 Annotation pour la désambiguïsation des termes [Kha00]

Cette approche consiste à utiliser une ontologie dans la phase d'indexation pour désambiguïser les termes extraits d'un document. Khan [Kha00] utilise un algorithme de désambiguïsation des concepts représentant un document pour affecter un terme dans un texte à un concept dans l'ontologie.

Les concepts non pertinents sont éliminés par la fixation d'un seuil basé sur le score de propagation est fixé. L'hypothèse de cette approche est qu'un terme dans un document pris isolément de son contexte peut être ambigu.

3.2.6 Annotation par ontologie de domaine [BL09]

Benyahia et al [BL09] ont proposé une approche d'annotation sémantique des pages web, l'idée de base est de représenter le document par un ensemble de mots clés, cet ensemble est le résultat de deux analyses, l'analyse statistique qui permet de sectionner les mots les plus fréquents dans le texte, et l'analyse sémantique permettant d'extraire les mots qui ont un poids sémantique. L'ensemble de ses mots sera relié ensuite à une ontologie de domaine par l'intervention de l'annotateur.

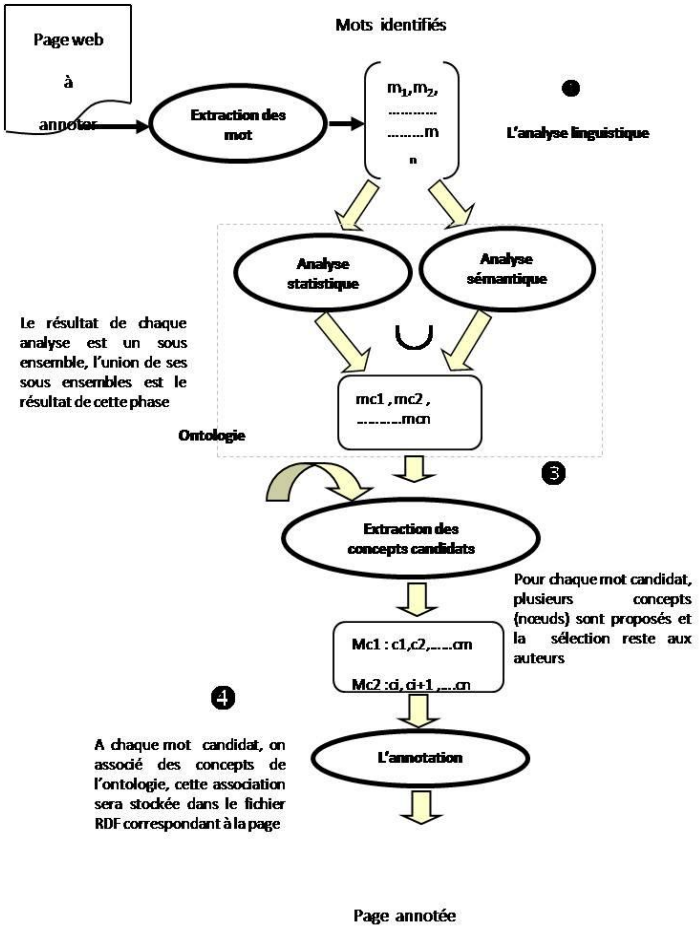

Figure 3.2 Processus d'annotation par ontologie de domaine [BL09]

3.2.7 MnM un outil basé sur l'ontologie [VM02]

Vargas-Vera et al. [VM02] a présenté le MnM, un outil d'annotation basé sur l'ontologie, qui intègre le navigateur Web, éditeur de l'ontologie et des API ouvertes pour fournir les deux supports automatiques et semi-automatiques pour l'annotation des textes dans les pages web. Il est capable d'extraire des informations à partir de pages web et les remplir dans un modèle prédéfini. En outre, une validation à base de type simple a été proposée afin de vérifier l'exactitude des contenus qui sont remplis dans le modèle.

3.2.8 La plate-forme KIM [PK03]

Popov et al. [PK03] ont développé une plate-forme de connaissance et de gestion de l'information (KIM) qui est basée sur une ontologie KIM et une base de connaissances massive pour annotation automatique des documents, indexation et recherche d'information. Selon l'hypothèse que les entités nommées (NE), tels que les personnes et l'emplacement qui sont désignés par leur nom, constituent la sémantique essentielle dans un document, L'annotation sémantique automatique est considérée comme le processus de reconnaissance et d'annotation des entités nommées(NE). Il fournit pour chaque entité nommée extraite (NE) avec deux types de liens: un lien vers la classe la plus spécifique dans l'ontologie KIM pour spécifier le type d'entités nommées et l'autre lien vers l'individu spécifique dans la base de connaissances.

3.2.9 Un framework pour l'annotation sémantique [ML11]

Ma et al. [ML11] ont proposé un cadre pour supporter le raisonnement sémantique du domaine et information linguistique qui sont incorporés dans les annotations de textes. Il utilise deux ontologies: (1) une ontologie de domaine pour fournir des étiquettes sémantiques (connaissances de domaine), et (2) une ontologie de la langue pour donner un modèle de texte (connaissances linguistiques). Dans le premier cas, comme il est représenté sur la figure 3.4, une assertion d'annotation sémantique est définie. Pour l'autre cas, il est représenté comme un ensemble d'axiomes OWL et SWRL [W3c04] des règles, ce qui contribue à combler les contraintes d'inférence.

Semantic Annotation is a tuple <**tf**,**ot**,**at**>
where
tf is a set of text fragment;
at is a set of semantic labels;
ot is a set of relations between **tf** and **ot**.

Figure 3.3 Le schéma d'annotation sémantique de Ma et al [ML11]

3.3 Annotation des documents par contexte

L'annotation des documents par contexte ne dépend pas seulement du contenu du document mais aussi du contexte. Nous présentons dans cette partie des travaux effectués dans l'axe d'annotation par contexte.

Un document pédagogique cite généralement d'autres documents comme références, la figure 3.4 illustre les différentes relations de citations entre les documents.

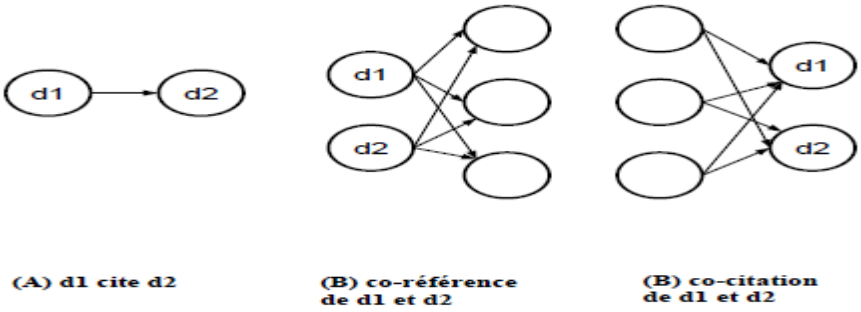

Figure 3.4 les relations entre les documents [Abr06]

Plusieurs cas de citations peuvent se présenter :

- Un document d1 référence un document d2.
- Les documents partagent une ou plusieurs références bibliographiques.

Nous présentons quelques travaux sur l'annotation par contexte :

1- La méthode de propagation de mots clés de Marchiori
2- Propagation de métadonnées de Prime
3- Annotation par le contexte de citation basée sur une ontologie

3.3.1 La méthode de propagation de mots clés de Marchiori [Mar98]

L'approche de Marchiori [Mar98] permet de propager des mots clés sur des pages Web.

Ces mots clés sont pondérés par un coefficient compris entre 0 et 1. Elle est basée sur l'hypothèse suivante :

Si une ressource P0 du Web a des métadonnées (mots clés) associées indiquant que le mot clé A a un poids v et s'il existe une ressource P00 dans le Web avec un hyperlien vers P0, alors les métadonnées de P0 sont propagées à P00. L'idée est que l'information contenue dans P0 est accessible par P00, étant donné qu'il existe un lien. [Abr06]

Marchiori[Mar98] a appliqué un facteur d'affaiblissement qui peut s'il est important supprimer des mots clés dans les références qui peuvent être importants et s'il est faible, la portée de la propagation est rapidement ingérable. [Abr06]

3.3.2 Propagation de métadonnées de Prime [Pri04]

Prime [Pri04] dans on approche affecte les métadonnées aux pages par la propagation dans le graphe du Web. Ces métadonnées représentent le type d'autorité, le type d'information et le type du site. Partant du principe de Marchiori[Mar98] , Prime se base sur l'hypothèse que si une page P contient un lien vers une autre page P0, alors ces pages partagent des métadonnées communes, mais l'approche ne s'applique pas sur tout le graphe du Web mais sur un graphe construit avec la méthode de co-citation définie par Prime.

3.3.3 Annotation par le contexte de citation basée sur une ontologie[Abr06]

Abrouk [Abr06] a présenté une approche et des outils pour l'annotation de documents en se basant sur des ontologies. Elle traite le problème d'annotation par une approche basée sur la relation de citation. Cette relation constitue la base d'une méthode pour affiner la propagation des annotations entre les documents. L'approche est indépendante du contenu et utilise un regroupement thématique des références construit à partir d'une classification floue non-supervisée. L'annotation étant basée sur l'utilisation d'ontologies, elle a également abordé le problème de l'enrichissement de l'ontologie afin de pouvoir prendre en compte les différentes évolutions des documents et affiner la phase d'annotation. Un outil, nommé RAS, Reference Annotation System, a été développé et des expérimentations ont été réalisées en utilisant la base Citeseer.

3.4 Discussion

Dans le Tableau (Tableau 3.1), une comparaison globale des travaux mentionnés ci-dessus est présentée. Les quatre premières colonnes répondent aux questions «Quoi», «Pourquoi» et "Comment" l'annotation sémantique est effectuée. Et puis, les deux autres colonnes sont utilisées pour décrire les facteurs d'exactitude et de stockage qui sont importants pour la formalisation d'une annotation sémantique. Chaque colonne dans ce tableau est présentée comme suit :

(1) La colonne «Domaines d'application» répond à la question "Que annoter?". Il décrit l'objet d'annotations sémantiques, qui repose sur le contexte de recherches (texte , vidéo, image et services web), dans notre cas on s'intéresse qu'aux travaux d'annotation des textes.

(2) La colonne «Façons d'annotation», répond à la question "Comment annoter?". Il décrit comment les annotations sémantiques sont faites. Le contenu de cette colonne peut être "Manuel", "semi-automatique" ou "automatique".

(3) La colonne "type ", répond à la question « comment la ressource est-elle traitée pour réaliser l'annotation ? », soit par contenu ou par contexte.

(4) La colonne "Ontologies " répond en partie à la question "Comment annoter?". Il décrit les ontologies qui sont utilisées dans la recherche correspondante.

(5) La colonne " vérification d'exactitude " décrit l'existence d'un mécanisme permettant de vérifier l'exactitude des annotations sémantiques existantes.

(6) La colonne "manière de stockage" décrit comment les annotations sémantiques sont attachées aux éléments annotés. Ils peuvent être intégrés dans la ressource à annoter (embarquées) ou stockée indépendamment de celle-ci (débarquées).

Cette comparaison met l'accent sur plusieurs informations clés dans les travaux de recherche sur les annotations sémantiques, y compris :

1-La plupart des recherches d'annotation sémantique sont axées sur l'utilisation des annotations sémantiques par contenu.

2- Le mécanisme utilisé est généralement semi-automatique.

3 Les mécanismes de vérification ne sont pas pris en compte par la majeure partie des recherches.

4-Les annotations sémantiques sont généralement intégrées à l'intérieur des objets d'annotation (embarquée).

Auterurs	Domaine d'application	Façon d'annotation	type d'annotation	L'ontologie utilisée	Vérification	stockage
Desmontils et al [DJ02]	texte	Semi-automatique	Par contenu	wordnet	Non	Embarqué
Njmogue et al [NF04]	texte	Semi-automatique	Par contenu	domaine	Non	Embarqué
Guarino et al [GM99]	texte	automatique	Par contenu	-Wordnet -L'ontologie Penman	Non	Embarqué
Baziz et al [Baz05]	texte	Semi-automatique	Par contenu	Ontologie lnguistique	Non	Embarqué
Khan et al [Kha00]	texte	Semi-automatique	Par contenu	domaine	Non	Embarqué
Vargas-Vera et al. [VM02]	texte	Manuelle, semi et automatique	Par contenu	domaine	Non	embarqué
Popov et al. [PK03]	texte	automatique	Par contenu	Kim	Non	embarqué
Ma et al. [ML11]	texte	automatique	Par contenu	domaine	Oui	débarqué
Benyahia et al [BL09]	texte	Semi-automatique	Par contenu	domaine	Non	débarqué
Marchiori et al [Mar98]	texte	automatique	Par contexte	Non spécifie	Non	débarqué
Prime et al [Pri04]	texte	automatique	Par contexte	Non spécifie	Non	débarqué
Abrouk et al [Abr06]	texte	automatique	Par contexte	domaine	Non	débarqué

Table 3.1: comparaison des travaux d'annotation sémantique

3.5 Conclusion

Dans ce chapitre, nous avons présenté une enquête sur un certain nombre de la littérature de l'annotation sémantique recueillies suivi par une comparaison détaillée et discussion.

Cette enquête montre que l'approche d'annotation s'effectue en passant par une analyse linguistique du texte qui peut être complétée par d'autres méthodes.

L'utilisation des ontologies augmente la représentativité des ressources documentaires, mais n'utiliser que le contenu de la ressource provoque un manque puisque :

- Une ressource documentaire peut ne pas avoir beaucoup de contenu. Par exemple le cas d'un document présentant une conférence va avoir de nombreuses références et peu de texte.
- Une ressource documentaire peut être représentée par un ensemble de métadonnées associées et donc le contenu n'est pas toujours disponible.

Donc, l'annotation par l'exploitation des références et citations (contexte) renforce sa représentation et sa sémantique.

Effectivement, notre travail rejoint dans sa deuxième partie celui d'Abrouk [Abr06] dans le sens ou on utilise aussi les liens afin de propager des annotations.

Pour les systèmes d'annotation, peu de travail avait été fait jusqu'alors quant au problème de leur validation. Il s'agit en effet de valider des systèmes qui avec des informations d'annotation, opérant des raisonnements cognitifs pour définir des informations requises.

A l'heure où les annotations pénètrent tout type de documents sur le web, comment admettre que les résultats d'une telle annotation qui va définir des ressources sur le web, ne soit rien d'autre qu'un acte de foi qu'un acte de confiance mutuelle entre les équipes d'annotations et les développeurs des systèmes de recherche qui y en prend en livraison.

Il peut arriver qu'une même ressource soit annotée par deux experts différents avec des manières différentes et donc cette ressource peut être annotée par des concepts d'ontologie disjoints.

Notre approche présente dans sa démarche un module de validation qui rend la base d'annotation cohérente, les différents modules de notre approche seront détaillés dans le chapitre qui suit.

Contributions

Dans ce chapitre, nous proposons une approche pour l'annotation semi-automatique des documents .Cette approche permet d'annoter un document avec contenu et avec contexte. Le processus d'annotation sera validé par un module de détection d'incohérence.

4.1Introduction

L'annotation d'un document consiste à le décrire par un ensemble des mots issue ou non d'un vocabulaire contrôlé. Dans notre contexte ces mots sont reliés par des relations sémantiques appartenant à l'ontologie du domaine. L'utilisation d'annotations de documents permet de décrire et d'utiliser au mieux les documents. Un document non annote est considéré comme inexploitable et donc impossible à le retrouver, Cette phase d'annotation est importante dans la recherche d'informations, car d'interrogation se base essentiellement sur la description des documents pour les retrouver.

A partir de ces constatations et du besoin des enseignants, nous avons développé une méthode d'annotation semi-automatique, l'annotateur (l'enseignant) restant décideur de la fiabilité de l'annotation.

Parti des constats : (1) qu'un document doit être bien décrit par des mots clés qui seront ensuite relié à des concepts d'ontologie afin de tirer profit du contenu, (2) qu'un document pédagogique référence généralement d'autres documents , (3) la base d'annotation peut contenir des redondances et des annotations d'une même ressource disjoints sémantiquement (puisque une ressources peut être annotée par des experts différents), nous présentons dans ce chapitre une approche pour l'annotation semi-automatique des documents pédagogique. Cette approche permet d'annoter une ressource par contenu et par contexte ; par contenu en basant sur la représentation du document par un ensemble des mots et relié ces mots à une ontologie. Par contexte, partant du principe qu'un document cite en générale autres documents, Nous propageons des annotations des documents cités sur le document citant.

4.2 Présentation de l'approche

Pour surmonter le problème du partage de document pédagogique et de faciliter l'accès au contenu, nous proposons une méthode d'annotation et de validation [KA16].La tâche de notre système est de prendre comme entrée un document Web et à mettre le même contenu enrichi par des annotations valides fondées sur des représentations de connaissances plus ou moins formelles.

Notre approche est composée de trois modules illustrés sur la figure 4.1.

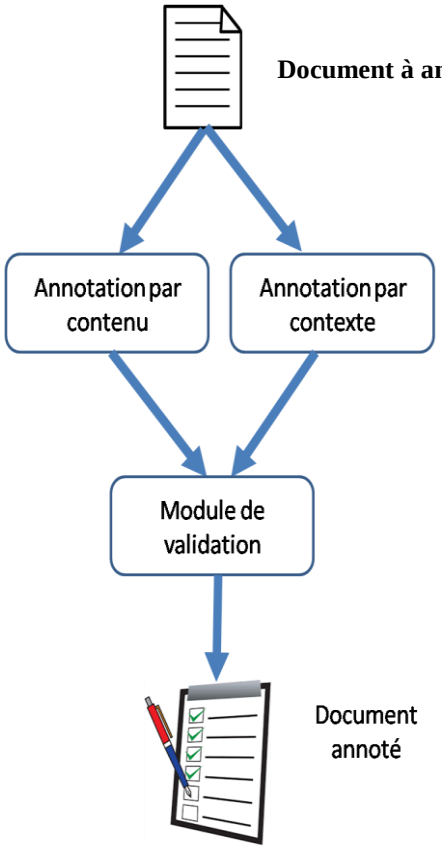

Figure 4.1. Schéma de l'approche proposée [KA16]

- Module d'annotation par le contenu: Le module qui extrait les mots clés candidats dans le document, dans cette phase deux types de calcul sont utilisés, la pondération des mots et le calcul de similarité. Les mots clés candidats sont ensuite combinés avec les concepts d'une ontologie de domaine.
- Module d'annotation par le contexte: partant du principe qu'un document pédagogique cite d'autres documents, le module va extraire les références citées dans le document et importer ses annotations comme annotation du document citant, on parle de la propagation des annotations.
- Module de validation: puisque notre base d'annotation sera construite de deux méthodes (par contenu et par contexte), une éventuelle redondance suspecte d'être existée, notre module de validation teste la cohérence des annotations, il se déclenche chaque fois qu'une annotation est créée, il élimine la redondance et prouver la cohérence de la base d'annotation.

Nous présentons ci-après le fonctionnement de chaque module

4.2.1 Module d'annotation par contenu

Les mots-clés, que nous définissons fournissent une représentation complète du contenu d'un document. Idéalement, les mots-clés représentent une forme inconduite du contenu essentiel d'un document.

Les mots-clés jouent un rôle crucial dans l'extraction des informations correctes selon les exigences des utilisateurs. Chaque jour, des milliers de livres, des articles sont publiés, ce qui rend très difficile de passer en revue tout le texte, mais il est nécessaire de disposer d'une bonne extraction d'information ou de méthodes de synthèse qui fournissent le contenu réel d'un document donné. En tant que tels, les mots clés efficaces sont une nécessité.

Puisque le mot-clé est la plus petite unité qui exprime la signification d'un document entier, de nombreuses applications peuvent en profiter comme l'Indexation automatique, le résumé automatique, la classification automatique, le regroupement, le filtrage automatique, la détection et le suivi des sujets, la visualisation de l'information et la définition des requêtes dans les systèmes de recherche d'informations (RI). Ils sont faciles à définir, à réviser, à rappeler et à partager. [KG10]

Plusieurs approches ont été proposées pour l'extraction automatique de mots clés, les approches statistiques qui sont simples et basées sur L'information statistique des mots pour identifier les mots clés dans le document. Les approches linguistiques utilisent les caractéristiques linguistiques des mots principalement des phrases et des documents, ils utilisent d'autres analyses comme l'analyse lexicale, l'analyse syntaxique et l'analyse du discours. [ZH08]

D'autres approches concernant l'extraction de mots-clés combinent principalement les méthodes mentionnées ci-dessus ou utilisent des connaissances heuristiques dans la tâche d'extraction de mots clés, telles que la position, la longueur, la disposition des mots, etc.

Notre module d'annotation par contenu est basé sur la recherche des mots clés qui représente mieux le document, une bonne sélection de ses mots offre aux annotateurs un large choix dans la phase d'annotation et donc une sémantique riche puisqu'ils seront associés aux concepts d'ontologie par le biais de l'annotateur. La figure 4.2 présente les phases de ce module

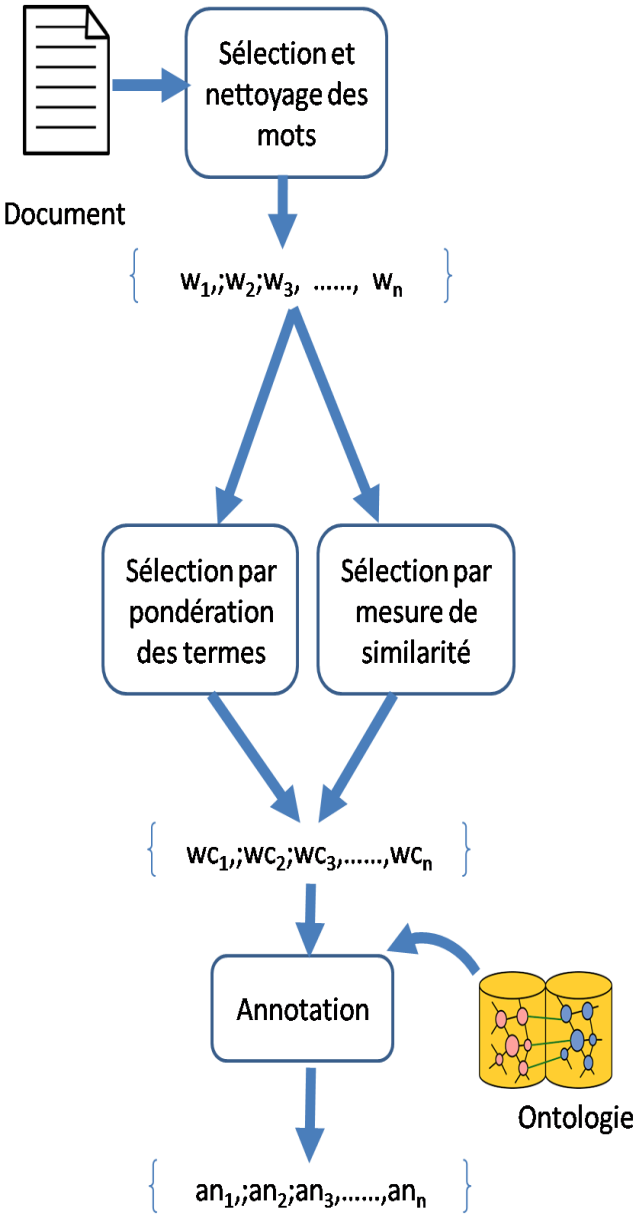

Figure 4.2 Annotation par contenu [KA16]

4.2.1.1 Sélection et nettoyage des mots

Un document est représenté par des mots-clés. L'indexation des textes comprend deux étapes : la recherche des mots caractérisant le contenu et l'évaluation du pouvoir de caractérisation de ces mots. Différents problèmes sont à résoudre :

- définir l'élément qui sera choisi comme unité d'indexation (radical, mot simple, groupe de mots) :
- choisir les mots représentatifs du document et ceux qui ne le sont pas, en fonction du contenu du document (mots d'indexation),
- évaluer le pouvoir de caractérisation de ces mots : certains mots sont plus importants que d'autres dans la caractérisation du contenu.

Au cours de ces différentes étapes, différents types de traitements sont appliqués qui peuvent être de type linguistique ou de type statistique.
La question principale est de savoir comment utiliser ces ressources textuelles accessibles à tous, comment tirer profit de ces bases de données linguistiques et comment extraire les mots qui seront utilisés par ce système d'annotation pour bien représenter le document. Le traitement linguistique représente le document à annoter par un ensemble de mots simples et importants.

Commençant par l'étape de segmentation du texte , cependant, lorsque l'on fait des statistiques sur les événements, nous voyons que les mots les plus fréquents sont des mots de fonction (ou mots des outils , des mots vides), comme par exemple en langue anglaise "of", "an", "the", etc. qui ne jouant qu'un rôle syntaxique et donnant peu de sens aux documents, de sorte qu'il ne serait pas nécessaire de les prendre en considération dans la phase d'annotation. L'élimination de ces mots vides est la deuxième étape.

4.2.1.2 Sélection par pondération des termes

L'objectif est maintenant de trouver les mots qui représentent mieux le contenu d'un document. Basant sur le principe de Luhn [Luh58] "quand un auteur écrit un texte, il répète certaines conditions pour développer un aspect du sujet", Il est généralement admis qu'un mot apparaît souvent dans le texte est un concept important. Ainsi, la première approche est de choisir les représentants des mots en fonction de leur fréquence d'apparition. Le plus simple est de fixer un seuil sur la fréquence : si une fréquence d'occurrence mot dépasse le seuil, il est considéré comme important pour le document. Mais, en général de la simple apparition de mot ne peut pas indiquer le sujet, le sens ou l'objet d'un texte.

Le processus de pondération devrait fournir une représentation iconique, compacte et informative du contenu du document. Il devrait fournir un indicateur important de distinguer les termes de chacun contre l'autre. Cet indicateur important (termes de poids) est souvent mesuré à partir de trois paramètres : la fréquence du terme, la fréquence de document du terme et la longueur standardisée du document. Plusieurs méthodes ont été proposées dans la littérature pour mesurer le terme «significatif». Nous sommes intéressés par la pondération locale dont le principe est le suivant :

La pondération locale mesure la représentation locale d'un terme. Elle prend en compte l'information locale du terme qui ne dépendent que du document donné, et donne l'importance du terme dans le présent document. Nous avons utilisé la fonction logarithmique qui combine TFij (la fréquence d'occurrence du terme ti dans le document dj) avec un logarithme, est donnée par :

$$imp = \alpha + \log(tf_{ij}) \qquad (1)$$

Où α est une constante.

Proposée par [GC88], a pour objectif d'atténuer les effets des différences importantes entre les fréquences d'occurrence des mots dans le document. Ainsi, en choisissant les mots qui ont des fréquences plus élevées que le seuil défini par l'utilisateur pour obtenir les mots dont l'informativité est la plus élevée.

4.2.1.3 Sélection par mesure de similarité

Dans cette étape, nous avons déterminé le poids d'un mot dans le document ; ce calcul de poids sémantique est basé sur la mesure de similarité.

Selon une étude de la littérature par [Tch12] sur les différentes mesures de similarité existantes, le principe de mesures fondées sur la distance taxonomiques est de compter le nombre d'arc séparant les deux sens dans une taxonomie. La figure (figure 4.3) [WP94] représente la relation entre deux concepts **c1** et **c2** dans une taxonomie par rapport à leurs concept commun le plus spécifique **c** et par rapport à la racine de la taxonomie.

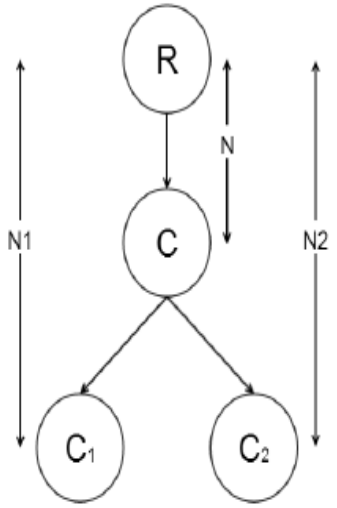

Figure 4.3 Les Distances utilisées par la mesure de similarité de Wu&Palmer[WP94]

Notons :

- C(C1, C2) est longueur du chemin le plus court en nombre d'arcs qui mène d'un concept C1 à un concept C2.
- P(C) est la longueur du plus court chemin entre le concept C et la racine.
- D (C1, C2) est le plus petit ancêtre commun des concepts C1 et C2, c'est le concept le plus spécifique qui les subsume.
- dis(C1/C2) est la distance en nombre d'arcs qui sépare C1 au D(C1, C2).

Wu et Palmer [WP94] ont défini une mesure de similarité entre concepts qui s'applique à un domaine conceptuel qui correspond à un point de vue donné pour lequel un mot a un seul sens et correspond donc à un seul concept. La similarité est définie à partir de la distance qui sépare deux concepts par rapport au concept le plus spécifique qui subsume les deux concepts dans le thesaurus, ainsi que la racine de la hiérarchie. La similarité entre C1 et C2 est (voir Figure 4.3):

$$sim_{\text{Wu et Palmer}}(C_1, C_2) = \frac{2*N3}{N1+N2+(2*N3)} \quad (1)$$

Plus formellement cette mesure revient à:

$$sim_{\text{Wu et Palmer}}(C_1, C_2) = \frac{2*P\ (D(C_1,C_2))}{P_c(C_1)+P_c\ (C_2)} \quad (2)$$

Rada [RM89] est le premier à utiliser la distance entre les nœuds correspondant aux deux sens sur les liens hyponym et hypernym:

$$\text{SimRada(c1 , c2)} = \text{d(c1 , c2)} = \text{N1} + \text{N2} \quad (3)$$

Les termes situés plus profondément dans la taxonomie sont toujours plus proches que les termes les plus généraux,

Lin [Lin98] propose une mesure de similarité qui calcule la proportion d'information commune entre deux concepts par rapport à leur description.

$$sim_{Lim}(C_1,\ C_2) = \frac{2.CI(D(C_1,C_2))}{CI(C_1)+CI(C_2)} \quad (4)$$

Leacock et Chodorow [LC98] ont également basé sur la mesure de la Rada, mais plutôt de normaliser la profondeur relative de la taxonomie en relation avec les sens, ils choisissent une normalisation par rapport à la profondeur totale de la taxonomie D et la normalise avec un logarithme.

$$\text{SimLCH} = -\log(\frac{\text{N1+N2}}{2*\text{D}}) \quad (5)$$

Dans un premier temps, Nous avons expérimenté plusieurs mesures de similarité sémantiques, nous avons ensuite choisi d'utiliser Wu & Palmer [WP94] en raison de sa meilleure performance vis-à-vis aux autres mesures de similarité sémantiques et sa simplicité offerte pour quantifier la similarité de deux concepts par la distance sémantique découverte par parcours de graphe.

Nous avons utilisé bibliothèque java (Java WordNet::Similarity[4]) existante qui implémente la plupart des mesures sémantiques existantes et manipule l'accès à WordNet 2.1

Dans cette phase, un mot est accepté si et seulement s'il est fortement lié à d'autres mots dans ce document. Cette décision dépend de la sélection d'un seuil défini par l'utilisateur appartient à l'intervalle [0,1].

4.2.1.4 Création de l'association

Dans cette étape on utilise une Ontologie de domaine, nous avons fait un passage des mots clés candidats à l'ontologie pour définir les concepts correspondants. A chaque passage d'un terme à l'ontologie, un ensemble de concepts sera présenté aux enseignants pour choisir les concepts à utiliser dans l'annotation. Une association sera créée, c'est l'annotation.

4.2.2 Module d'annotation par contexte

Nous avons basé sur l'algorithme de abrouk[Abr06] , Le choix de cet algorithme est dû à la similarité avec notre approche dans le sens où on utilise les liens de citations afin d'annoter les documents. Ceci revient à une propagation d'annotations.

Un document pédagogique fait référence généralement à autres documents. Dans cette phase, nous nous intéressons à la partie des références dans le document pour importer les annotations des documents cités par un document citant d. La figure 4.4 schématise le module d'annotation par contexte.

La phase de propagation passe par les étapes suivantes :

1. Récupérer l'ensemble des documents cités par d
2. Sélectionner les annotations pour chaque document cité.
3. Importer les annotations des documents cités par d.
4. Ajouter les annotations importées aux annotations du document d.

Pour chaque référence de la phase précédente, Nous importons leurs annotations qui sont généralement des concepts définis dans l'ontologie utilisée pour l'annotation sans avoir besoin du contenu du document. Cet ensemble d'annotation sera considéré comme des annotations du document D.

Nous définissons l'ensemble des annotations importées pour le document d comme suit : $$Annot_d = \bigcup_{df \in Refd} Annot(df) \qquad (6)$$

Annot(df) regroupe l'ensemble des annotations du document df

[4]http://www.cogs.susx.ac.uk/users/drh21/

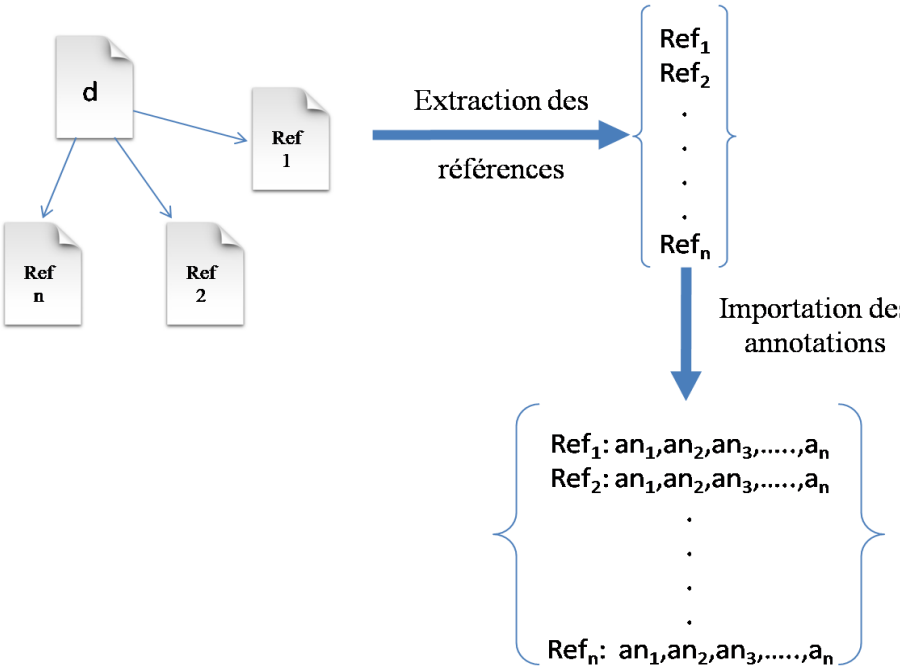

Figure 4.4 L'annotation par contexte

Le processus d'annotation par contexte est décrit par l'algorithme suivant :

Algorithme 1 : Annotation_par_contexte

refD : L'ensemble des référence de D

$Annot_D$: les annotations du document D

Annot(df) : les annotation du document cité en référence

Debut

refD← {ref1 ,ref2 ,…..,refn}//

Pour df∈refD//*pour chaque document appartient à l'ensemble des références citées dansD*

$Annot_D$←annot D + annot(df) // *importation des annotations*

Fin.

4.2.3 Le module de validation

Après l'étude de l'état de l'art du chapitre 3, nous avons constaté que pour les systèmes d'annotation rien ou presque n'avait étais fait quant au problème de leur validation, surtout quand il s'agit de valider des annotations sémantiques qui sont défini d'une manière souvent approximative.

Partant des deux probabilités suivantes :

1 : La probabilité que deux enseignants différents peuvent choisir le même concept (mots-clés) pour décrire un mot est faible, ce qui rend difficile l'obtention d'une cohérence dans l'annotation.

2: Les éléments de l'annotation sont des concepts définis dans l'ontologie globale d'annotation, c'est à dire un ensemble avec une redondance des éléments, car deux références peuvent avoir des concepts communs.

Nous présentons un module de validation des annotations créées et importées, la figure 4.5 montre les étapes de validation.

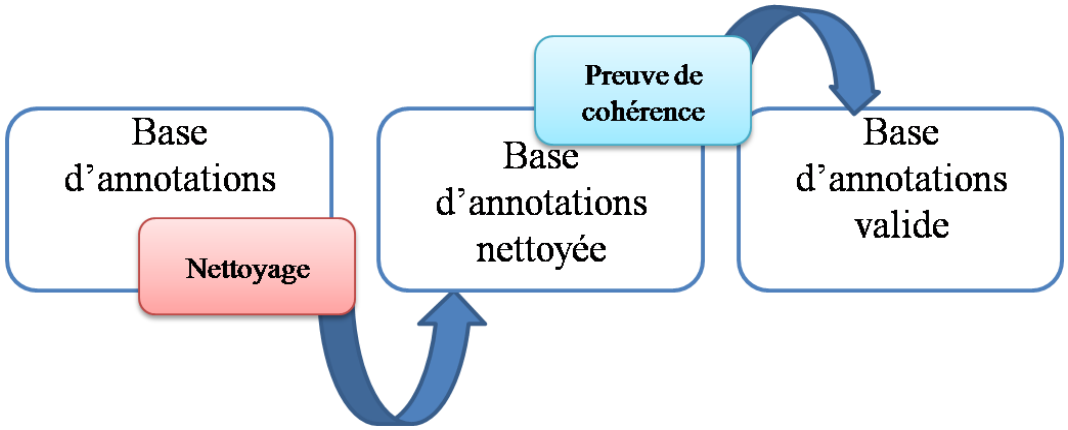

Figure 4.5 La validation des annotations

Les spécifications de l'incohérence portant sur :

1-La redondance des annotations ; et donc la nécessité de nettoyage

2-Des annotations en conflit qui rendre l'ensemble des annotations incohérent d'où la preuve de cohérence devient une étape très utile.

4.2.3.1 Le nettoyage

Partant de principe que deux références peuvent avoir des concepts communs, l'élimination de cette redondance est la première phase dans le module de validation.

L'idée de base consiste à parcourir la base d'annotation en cherchant les annotations en redondances, une fois une redondance est détectée le processus de suppression est déclenché en gardant une seule annotation.

4.2.3.2 Preuve de cohérence

Appliquer un mécanisme de validation, pour rendre l'ensemble des annotations cohérent, il peut arriver qu'une ressource sera annoté par des concepts disjoints sémantiquement et donc on doit recalculer la sémantique afin de garder un et éliminer les autres. Nous avons défini un mécanisme de recherche d'incohérence, ce mécanisme permettent de focaliser la recherche d'incohérence à chaque fois l'annotation est créée. L'annotation doit être révisée une fois les incohérences sont détectées.

L'idée de base consiste à décomposer la base des annotations en sous-ensembles maximaux des annotations portant sur les mêmes ressources, une fois cette décomposition

effectuée, la preuve de cohérence se ramène à des tests simples de cohérence pour chaque sous-ensemble, la détection des annotations en conflit d'une même ressource rend le sous-ensemble incohérent et les annotations d'un sous-ensemble ne seront jugées « correctes » qu'après avoir prouvé la cohérence de sous-ensemble. La validation de l'ensemble des annotations consiste à rendre chaque sous-ensemble cohérent.

Nous avons appliqué la mesure de Wu&Palmer (citée dans la section 4.2.1.3**)** pour prouver la cohérence ou détecter l'incohérence d'un sous ensemble.

Nous définissons un sous ensemble Sn1 des annotations d'un mot m1 par :

Sn1= {(c1,c2,…,cn) | m1 est annoté par c1 , c2,….,cn}

Soit un mot m1 de sous ensemble Sn1 annoté par deux concepts C1 et C2. La mesure de similarité entre C1 et C2 peut être catégorisée en deux cas :

Catégorie	signification	décision
C1 et C2 sont équivalents	Annotation correcte	Sn1 cohérent
C1 et C2 sont disjoints sémantiquement	Entre C1 et C2, l'un est incorrect	Sn1 incohérent

Tableau 4.1 Les résultats possibles de conflit entre 2 concepts d'annotation

Si les mesures de similarité de chaque couple des concepts est supérieur à un seuil défini par l'annotateur alors le sous ensemble est considéré comme cohérent. Le cas contraire, un état d'incohérence sera déclaré.

Une fois l'incohérence est détectée, le processus de validation déclenchera afin de rendre le sous ensemble cohérent dont le principe est le suivant :

Nous appliquons une mesure de similarité entre (m1, c1), et (m1, c2) par la mesure de Wu&Palmer (citée dans la section 4.2.1.3**)**, la décision de la suppression ou la conservation d'annotation reste à l'annotateur.

Aucune suppression d'une annotation n'est autorisée qu'après la confirmation de l'annotateur. Notre module de validation aide à la prise de la décision sur une telle annotation incohérente. Cette phase reste donc semi-automatique.

4.3 Conclusion

Ce chapitre décrit notre approche d'annotation de documents. L'annotation manuelle de document est une tâche difficile, voire impossible à réaliser, compte tenu du temps que cela nécessite. Au cours de ce chapitre, nous avons présenté les fondements théoriques d'une

nouvelle approche d'annotation sémantique valide du document basé sur deux méthodes d'annotation, par contenu et par contexte suivi d'une phase de validation.

Nos contributions ont porté sur 3 aspects : (1) l'annotation par contexte, qui propage les annotations des références citées dans un document sur le document citant sans avoir besoin de son contenu. Et(2) par contenu qui applique des calculs pour extraire les mots clés représentant le document afin de les relier à une ontologie de domaine, ce qui améliorera le processus de recherche de documents et résoudra le problème de multilinguisme puisqu'un terme exprimé dans différentes langues est associé à un seul concept. Un module de validation est appliqué à la fin de chaque phase d'annotation pour garder la base d'annotation cohérente. Notre approche n'est pas restée au niveau théorique, elle a été implémentée et le résultat de l'annotation a été testé sur un corpus de documents et les évaluations sont très encourageantes. L´expérimentation et l´évaluation de l'approche sont présentées dans le chapitre qui suit.

Implémentation et évaluation

La mise en œuvre de l'approche proposée a été suivie par une série d'expérimentations.

Dans ce chapitre, Nous décrivons les expérimentations que nous avons menées et les résultats obtenus.

5.1 Introduction

Nous avons présenté un système qui permet l'annotation sémantique d'un document (cf chapitre 4). Ce chapitre a pour objectif de l'évaluer et de tester ses performances. Nous souhaitons, en utilisant un ensemble de documents pédagogique, illustrer les points suivants : (1) les avantages d'utilisation des deux techniques d'annotation, par contenu et par contexte, (2) l'impact de d'application de module de validation pour rendre la base d'annotation cohérente.

Dans ce chapitre, nous présentons plus particulièrement les trois expérimentations réalisées en utilisant notre système. La première série d'expérimentations décrit les tests de mesure de l'efficacité de notre approche d'annotation mixte par contenu et par contexte.

Une deuxième série d'expérimentations se base sur le test d'approche de point de vue l'efficacité de notre module de validation. Enfin, la troisième série d'expérimentations porte sur la comparaison de notre système avec autres travaux de recherche.

Nous terminons ce chapitre avec une discussion

5.2 L'environnement de développement

Au cours de notre phase d'expérimentation, nous avons travaillé sur une machine qui a les caractéristiques suivantes :

-Un processeur I5

- RAM de 4Go.

5.3 Le corpus de teste

Notre corpus de test, nous avons recueilli un ensemble de 160 documents de domaine informatique constitués de 53 cours, 40 travaux dirigés, 47 Présentations PowerPoint et 20 travaux pratiques. La longueur moyenne de ces documents dans le corpus est de 10 pages.

5.4 Les ontologies utilisées

Nous avons utilisé plusieurs ontologie dans les différentes phases d'expérimentations DMOZ est la plus vaste et qui englobe les différents notions qui concerne notre domaine d'application « informatique ». DMOZ[5], aussi connu comme L'Open Directory Project (ODP), est le plus grand et le plus complet des répertoires du Web édités par des êtres humains. Il est développé et maintenu par une vaste communauté mondiale d'éditeurs bénévoles. Pour annoter les documents nous avions besoin d'une ontologie que nous avons limitée au domaine de l'informatique (Computer).

[5] http://dmoztools.net/

5.6 Langage de représentation d'annotation

Nous avons stockés nos annotations sous format de triplet RDF, l'avantage que l'on obtient est que l'information correspond directement et sans ambiguïté à un modèle, un modèle qui est décentralisé, et pour lequel il existe de nombreux génériques parsers déjà disponibles. Les données RDF devenus une partie du Web sémantique, donc les avantages de la rédaction de nos annotations dans RDF dessine maintenant des parallèles avec la rédaction des informations en HTML dans les premiers jours du Web.

5.7 L'évaluation

Pour évaluer le processus d'annotation, le corpus a été annoté par deux experts pour chaque document pédagogique, ils spécifient son type.

5.7.1 L'évaluation des modules d'annotations (par contenu et par contexte)

Pour évaluer les performances de fusion des deux modules d'annotations, nous avons construit, en premier lieu, une collection de test des modules d'annotation de notre système indépendamment, par contenu et par contexte en termes de nombre d'annotations créées. Le tableau 5.1 présente les résultats d'expérimentation.

Nature de Document	N	Nctn	Nctx
Cours	3650	2135	1515
Travaux dirigés	700	650	50
Travaux pratiques	120	112	8
Presentation power point	790	425	365

Tableau 5.1 : Résultats de comparaison des types d'annotations

Nous définissons, *N* : nombre total d'annotations

Nctn : nombre d'annotations créées par le processus par contenu

Nctx : nombre d'annotations créées par le processus par contexte

Discussion

Nous remarquons (figure 5.1) que pour les cours et les présentations power point, le nombre d'annotations créées en appliquant le module d'annotation par contexte est de 48% et ceci dû à la richesse des cours et des présentations par les références qui font l'objet de ce type d'annotation. Par contre pour les travaux dirigés et les travaux pratiques, l'annotation par

contexte n'a pas dépassé les 7% du nombre des annotations et ce taux est expliqué par la nature des travaux dirigés et travaux pratiques qui généralement ne citent pas trop de documents.

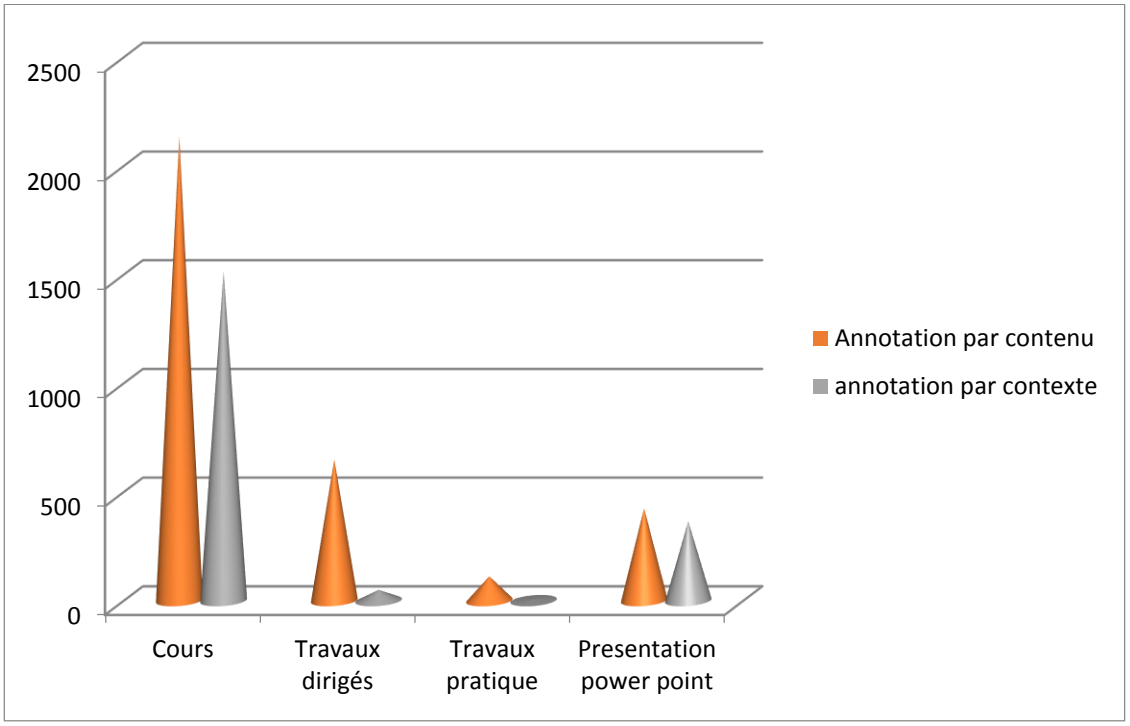

Figure 5.1 : Test sur le type d'annotation

Nous notons que l'utilisation des deux types d'annotations (selon le contenu et le contexte) renforce la sémantique des documents ; notre outil fusionne les annotations créées par l'extraction de mots significatifs du document et ceux hérités par la phase d'annotations par propagation les ressources citées dans le document annoté, ce qui augmente le nombre d'annotations créées.

5.7.2 L'évaluation de l'intégration du module de validation

Une deuxième série d'expérimentations a été menée dans l'objectif de montrer l'importance de notre module de validation dans la phase d'annotation ; Tout d'abord, nous avons testé le système sans l'application du module de validation ; les résultats du processus d'annotation effectué par le système sont détaillées dans le tableau 2.

L'évaluation s'est basée, particulièrement, sur un indice de qualité Iqa défini comme suit :

$$I_{qa} = \frac{Nac}{Na} \qquad (1)$$

Na: Nombre d'annotations.

Nac: Nombre d'annotations correctes

Nature de Document	Na	Nac	Iqa(%)
Cours	3650	2800	77%
Travaux dirigés	700	480	69%
Travaux pratiques	120	80	67%
Presentation power point	790	580	73%

Tableau 5.2 : résultats d'annotations sans l'application du module de validation

En deuxième étape, nous avons testé le système avec l'application du module de validation, le tableau 5.3, montre les résultats obtenus.

Nature de Document	Na	Nac	Iqa(%)
Cours	2850	2800	98%
Travaux dirigés	500	480	96%
Travaux pratique	87	80	92%
Presentation power point	600	580	97%

Tableau5.3 : résultats d'annotations avec l'application du module de validation

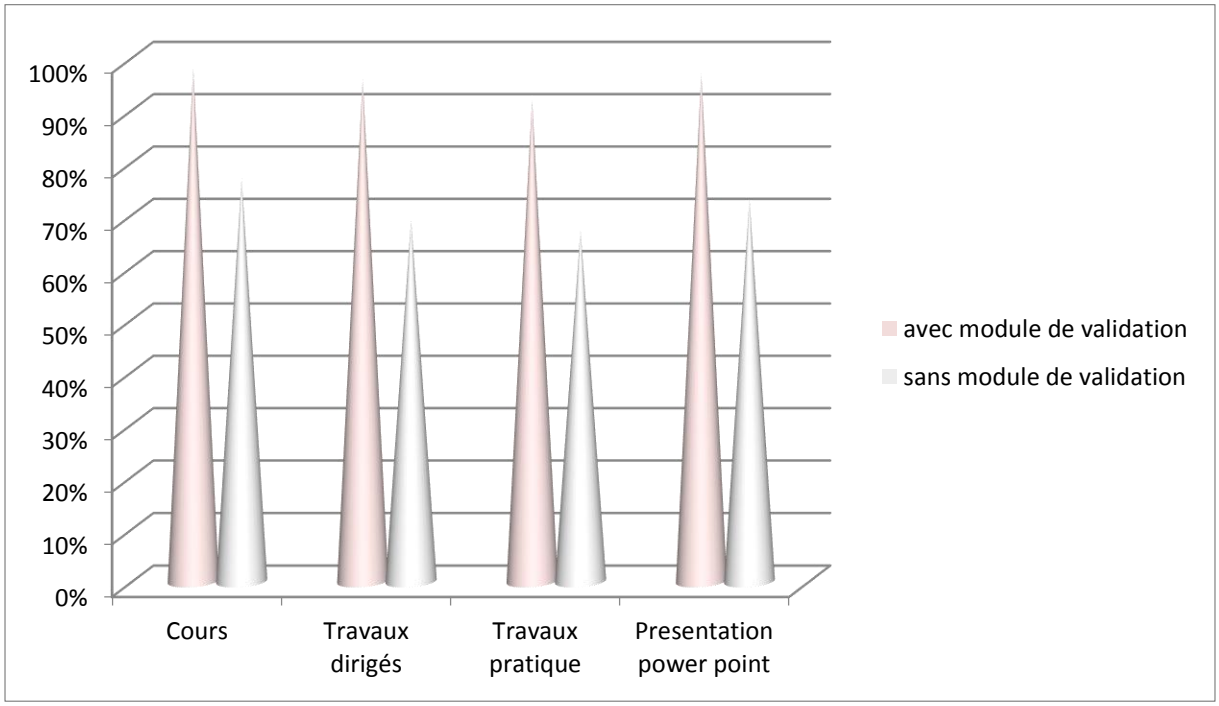

Figure 5.2 : l'index de qualité

Discussion

Nous notons que l'annotation sémantique des documents pédagogiques est étroitement liée à l'intégration de la phase de validation dans l'annotation sémantique (figure 5.2). L'indice de qualité "IQA" de l'annotation en intégrant le module de validation est supérieur à celle de l'annotation qui n'utilise le module de validation (tableau 5.2, tableau 5.3). Ceci est expliqué par le fait que le nombre d'annotation créée est réduit (3650 sans module de validation et 2850 avec le module de validation pour cours) qui augmente la "IQA", puisque le module de validation nettoie et élimine les annotations incohérentes et redondantes. L'ajout d'une nouvelle annotation peut rendre la base d'annotations incohérente et donc la révision est nécessaire afin de la rendre cohérente soit par la preuve de la cohérence ou l'élimination des incohérences surtout après l'étape de fusion des deux ensembles d'annotations (par le contenu et par le contexte) où la redondance peut se produire.

Nous notons également que le nombre d'annotations correctes est important (2800 pour les cours) et cela est justifié par l'utilisation de deux types de sélection de mots clés qui représente le document dans le module d'annotation par le contenu, la sélection par la pondération et la sélection par mesure de similarité qui offre aux annotateurs plus de mots candidats dans la phase d'annotation.

Ce dernier est lui-même dépend du type de document annoté qui est un facteur important affectant la qualité de l'annotation, le document de cours a atteint 98%, par contre pour les travaux dirigés et travaux pratiques n'a pas dépassé 92%. Cela se justifie par la richesse du contenu du cours et puisqu'ils citent plusieurs documents comme références qui augmentent le nombre d'annotations.

5.7.3 Comparaison avec autres approches

Pour bien testé les performances de notre approche, Nous avons procédé à la comparaison de notre approche avec celle de Benyahia et al [BL09] (cf **3.2.6)** qui est basée sur l'annotation par contenu à l'aide d'une ontologie de domaine. Le tableau 5.4 montre les résultats obtenus :

L'approche	**Iqa**
Notre Approche [KA16]	93%
L'approche de Benyahia et al [BL09]	78%

Tableau 5.4 : résultats de comparaison avec l'approche de Benyahia

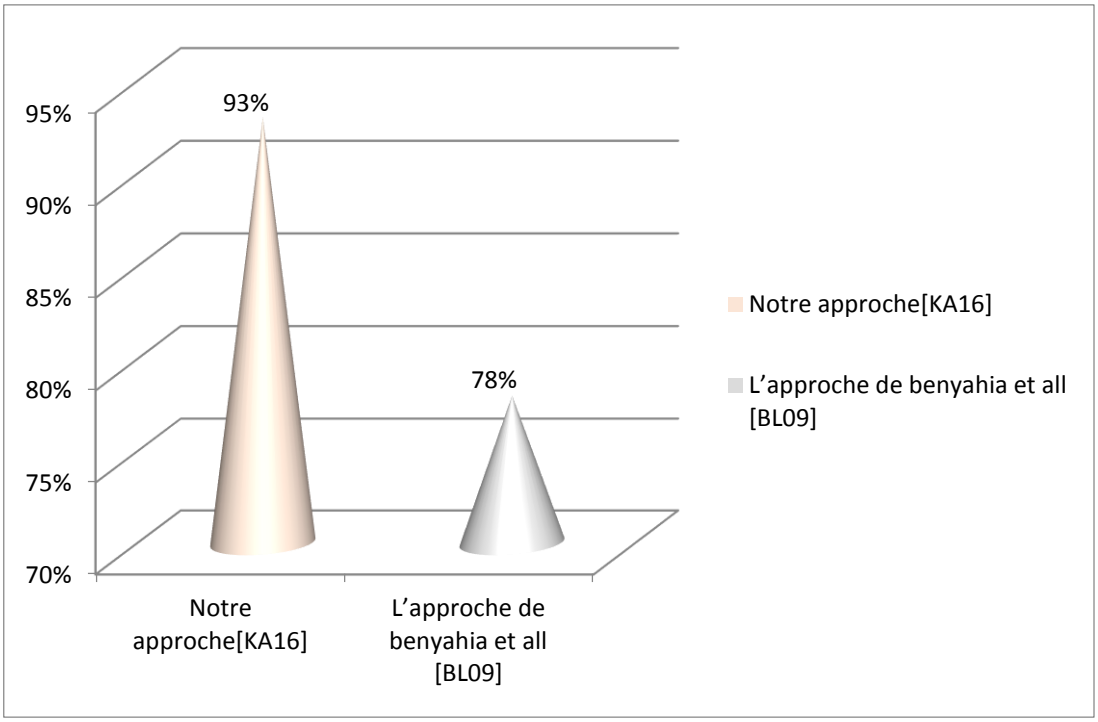

Figure 5.3 : Résultat de comparaison

Discussion

Les résultats obtenus sont encourageants, ceci est expliqué par le fait que l'approche de Benyahia et al [BL09] est basée sur le contenu du document, les annotations obtenues ne sont pas validées ce qui créée des incohérences. Par contre notre approche enrichie les documents sur les deux volets, contenu et contexte ce qui augmente le nombre d'annotation, notre module de validation élimine les incohérences dans la base d'annotation et donc fournie aux moteurs de recherche des ressources bien enrichi sémantiquement.

5.7 Conclusion

Dans ce chapitre, nous avons présenté l'expérimentation que nous avons menée au cours de notre travail en décrivant en premier lieu notre protocole d'expérimentation. D'un point de vue pratique, nous avons implémenté cette approche et procédé ensuite à l'évaluation sur un corpus de documents. Nous avons présenté trois phases d'évaluation, la première concerne l'utilisation des deux techniques d'annotation par contenu et par contexte, la deuxième concerne l'importance d'intégration du module de validation dans la phase d'annotation. La comparaison avec d'autres annotations effectuées par les auteurs des documents nous a permis de bien évalué notre approche. Les résultats obtenus sont très encourageants, 95% des annotations sont correctes ce qui montre l'importance de la validation dans la phase d'annotation et la fusion des deux techniques par contenu et par contexte.

Conclusion Générale

Nous venons de présenter dans cette thèse le fruit de nos travaux de recherche portant sur « l'annotation sémantique des documents ». Notre travail s'appuie sur les théories, les méthodes et les techniques développées en Web Sémantique.

En effet, le Web Sémantique a besoin pour son développement futur de pouvoir créer des annotations à partir de représentations formelles de la connaissance d'un domaine, telles que les ontologies. Une brève présentation du web sémantique ainsi que ses langages a été présentée dans le premier chapitre.

Nous avons présenté dans le second chapitre les annotations sémantique, un état de l'art qui décrit les différentes approches proposées surtout ceux qui concernent l'annotation par le contenu des documents et les annotations par contexte indépendamment du contenu, suivi par un tableau récapitulatif et comparatif des différents travaux abordés font l'objet du troisième chapitre.

Partant du résultat de ces études, dans un quatrième chapitre, nous avons présenté une approche d'annotation sémantique et de validation basée sur les deux types d'annotation par contexte et par contenu. Notre approche dans sa phase d'annotation par contenu vise à représenter le document par un ensemble des mots clés, cette représentation est le résultat de deux calculs, la pondération des termes et le calcule de similarité, ces mots clés seront ensuite relié à une ontologie de domaine. L'annotation par contexte et partant du principe qu'un document pédagogique cite généralement d'autres documents, nous avons propagé les annotations des documents cités sur le document citant. L'ajout des annotations soit par héritage ou par création peut rendre la base d'annotation incohérente. Un module de validation est la phase finale de notre approche qui sert à rendre la base cohérente soit par l'élimination de redondance ou par l'élimination de cohérence.

L'approche d'annotation a été implémentée pour ensuite être validée. Nous mettons en valeur l'intérêt de notre approche par des expérimentations appliquées sur une base de documents pédagogiques. Les résultats obtenus ont montré d'un côté l'impact de la méthode d'annotation hybride par contexte et annotation par contenu qui enrichit la sémantique des documents et d'autre coté l'importance de la validation dans la phase d'annotation. Ce qui offre un acte de confiance mutuelle entre les annotateurs et les développeurs des moteurs de recherche qui en prend livraison, surtout quand on sait que les annotations sont et seront amenés de plus en plus souvent à aider les moteurs de recherche à la prise de décisions pour mieux répondre aux requêtes des enseignants.

Il est important de noter ici que l'opérationnalisation des processus d'annotation sémantique et de validation est évidemment indissociable d'une réflexion complète sur l'utilisation de ces processus dans un contexte particulier.

En résumé, on peut constater que nos travaux apportent un ensemble de propositions non seulement pour la réalisation des activités d'annotation sémantique des documents mais aussi pour leur utilisation dans la maintenance des bases d'annotation.

Nos perspectives consiste à enrichir notre approche avec de nouvelles fonctionnalités comme :

- L'annotation sémantique de contenu multimédia.
- L'amélioration des interfaces hommes machines, afin de les rendre plus interactives et plus dynamiques avec les différents objets manipulés.
-Appliquer notre module de validation sur des bases d'annotations volumineux afin de tester sa robustesse.

Références bibliographiques

[**Abr06**] Lylia Abrouk. Annotation de documents par le contexte de citation basée sur une ontologie. Interface homme-machine [cs.HC]. Université Montpellier II - Sciences et Techniques du Languedoc, 2006.

[**BA04**] Bourigault, D., Aussenac-Gilles, N., & Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes Un cadre unificateur pour trois études de cas. Revue d'Intelligence Artificielle, 18(1), 87-110.

[**Baz05**] M. Baziz. Indexation conceptuelle guidée par ontologie pour la recherche d'information. PhD thesis, Institut de recherche en informatique de Toulouse, université Paul Sabatier, 2005.

[**BC02**] S. Bechhofer, L. Carr, C. Goble, S. Kampa, and T. Miles-Board, "The Semantics of Semantic Annotation," in Proceedings of the 1st International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE), Irvine, California, 2002, vol. 2519, pp. 1152–1167.

[**BL09**] Benyahia, K., Lehireche, A., & Latreche, A. (2009). Annotation Sémantique De Pages Web. In CIIA.

[**BP10**] BOURHIS .A , PERRIN.N (2010). Etat de l'art sur les Technologies du Web sémantique

[**CB04**] Charlet, J., Bachimont, B., & Troncy, R. (2004). Ontologies pour le web sémantique. Revue I3, page 31p.

[**Cha02**] Charlet, J. (2002). L'ingénierie des connaissances: développements, résultats et perspectives pour la gestion des connaissances médicales.

[**Cor06**] Corcho, O. (2006). Ontology based document annotation: trends and open research problems. International Journal of Metadata, Semantics and Ontologies, 1(1), 47-57.

[Dav10] Davidson, P. (2010). Designing uri sets for the uk public sector. UK Chief Technology Officer Council.

[**DJ02**] E. Desmontils and C. Jacquin. Indexing a web site with a terminology oriented ontology. The Emerging Semantic Web, I.F. Cruz, S. Decker, J. Euzenat and D. L. McGuinness Ed, pages 181–197, 2002.

[**DJM02**] E. Desmontils, C. Jacquin, and E. Morin. Indexation sémantique de documents sur le web : application aux ressources humaines. In Proceedings of Journées de l'AS-CNRS Web sémantique, Octobre 2002.

[**DO03**] Daconta, M. C., Obrst, L. J., & Smith, K. T. (2003). The semantic web: a guide to the future of XML, web services, and knowledge management. John Wiley & Sons.

[**EM92**] C. Enguehard, P. Malvache, and P. Trigano. Indexation de textes : l'apprentissage des concepts. In Proceedings of 14th conference on Computational linguistics, pages 1197–1202, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[**Flo07**] Florence Amardeilh. Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle. domain stic.gest. Université de Nanterre - Paris X, 200

[**GC88**] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513–523, 1988.

[**GM99**] Guarino, N., Masolo, C., & Vetere, G. (1999). Ontoseek: Content-based access to the web. Intelligent Systems and Their Applications, IEEE, 14(3), 70-80.

[**Gru93**] Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge acquisition, 5(2), 199-220.

[**Gua98**] Guarino, N. ed. 1998. Formal Ontology in Information Systems. Amsterdam: IOS Press. Proceedings of the First International Conference (FOIS '98), June 6–8, Trent, Italy.

[**Hab05**] HABERT B., Instruments et ressources électroniques pour le français, Collection "L'essentiel Français", Ophrys, Paris, 2005, 169 p

[**Han05**] Handschuh, S. (2005). Creating ontology-based metadata by annotation for the semantic web (Doctoral dissertation, Karlsruhe, Univ., Diss., 2005).

[**Her05**] Hernandez, N. (2005). Ontologies de domaine pour la modélisation du contexte en recherche d'information (Doctoral dissertation, Université Paul Sabatier).

[**JS99**] Jones, S. et Staveley, M. (1999). Phrasier : a System for Interactive Document Retrieval Using Keyphrases. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

[**KA16**] Kadda, B., & Ahmed, L. (2016). Semantic Annotation of Pedagogic Documents. International Journal of Modern Education & Computer Science, 8(6)

[KG10] Kaur, J., & Gupta, V. (2010). Effective approaches for extraction of keywords. Journal of Computer Science, 7(6), 144-148.

[**Kha00**] Khan, L. R. (2000). Ontology-based information selection (Doctoral dissertation, University of Southern California).

[**KM10**] KIM, S. N., MEDELYAN, O., KAN, M. et BALDWIN, T. (2010). Semeval-2010 Task 5 : Automatic Keyphrase Extraction from Scientific Articles. In Proceedings of the 5th International Workshop on Semantic Evaluation.

[**KP04**] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2. pp. 49–79, 2004.

[**LC98**] LEACOCK, C. et CHODOROW, M. (1998). Combining local context and wordnet similarity for word sense identification. WordNet : An Electronic Lexical Database. C. Fellbaum. Ed. MIT Press. Cambridge. MA

[**Lin08**] Y. Lin, "Semantic Annotation for Process Models: Facilitating Process Knowledge Man-agement via Semantic Interoperability," PhD Thesis, Norwegian University of Science and Technology, 2008.

[Lin98] Lin, D. (1998, July). An information-theoretic definition of similarity. In ICML(Vol. 98, pp. 296-304).

[**LL08**] Litvak, M., & Last, M. (2008, August). Graph-based keyword extraction for single-document summarization. In Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization (pp. 17-24). Association for Computational Linguistics.

[**Luh58**] H. Luhn. The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2) :159–165 and 317, April 1958.

[**Mar98**] M. Marchiori. The limits of web metadata, and beyond. In Proceedings of Seventh International World Wide Web Conference, pages 1–9, Australia, 1998.

[**ML11**] Y. Ma, F. Lévy, and S. Ghimire, "Reasoning with Annotations of Texts," in Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference, Palm Beach, Florida, USA, 2011, pp. 192–197.

[**MW08**] Medelyan, O. et Witten, I. (2008). Domain-Independent Automatic Keyphrase Indexing with Small Training Sets.

[**NF04**] W. Njmogue, D. Fontaine, and P. Fontaine. Identification des thèmes d'un document relativement à un référentiel métier. In Proceedings of MAJECSTIC'04, 13-15 Octobre 2004.

[**NFF04**] W. Njmogue, D. Fontaine, and P. Fontaine. Indexation des documents dans un référentiel métier. In Proceedings of Workshop ALCAA 2004, Agents Logiciels - Coopération Apprentissage - Activité humaine, 7-18 Juin 2004.

[**OM06**] E. Oren, K. H. Möller, S. Scerri, S. Handschuh, and M. Sintek, "What Are Semantic Annotations," Technical Report, Galway, Ireland, 2006.

[**PG04**] Prié, Y., & Garlatti, S. (2004). Méta-données et annotations dans le Web sémantique. Revue I3 Information-Interaction-Intelligence, 4, 45-68.

[**PK03**] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov, "KIM – Semantic Annotation Platform," in Proceedings of the 2nd International Semantic Web Conference, Sanibel Island, FL, USA, 2003, vol. 2870, pp. 834–849.

[**Pri04**] C. Prime-Claverie. Vers une prise en compte de plusieurs aspects des besoins d'information dans les mod`eles de la recherche documentaire : Propagation de métadonnées sur le World Wide Web. PhD thesis, Ecole supérieure des Mines de Saint-Etienne, 2004.

[**PZ12**] Paroubek, P., Zweigenbaum, P., FOREST, D. et GROUIN, C. (2012). Indexation Libre et Contrôlée d'Articles Scientifiques Présentation et Résultats du Défi Fouille de Textes DEFT2012.

[**Rij79**] V. Rijsbergen. Information Retrieval, 2nd edition. Dept. of Computer Science, University of Glasgow, London, 1979.

[**RM89**] RADA, R., MILI, H., BICKNELL, E. et BLETTNER, M. (1989). Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1):17–30.

[**Sal69**] Salton, G. (1969). A comparison between manual and automatic indexing methods. American Documentation, 20(1), 61-71.

[**Sal86**] G. Salton. Another look at automatic text-retrieval systems. Commun. ACM, 29(7) :648–656, 1986.

[**SB98**] Studer R., Benjamins V.R. & Fensel D., Knowledge engineering: principles and methods, in IEEE Transactions on Data and Knowledge Engineering, 25(1&2), 1998, pp.161-197.

[**Sma73**] H.G. Small. Co-citation in the scientific literature. Society for Information Science, 24, 1973.

[**TA09**] H. N. Talantikite, D. Aissani, and N. Boudjlida, "Semantic annotations for web services discovery and composition," Comput. Stand. Interfaces, vol. 31, no. 6, pp. 1108–1117, Nov. 2009.

[**Tch12**] Tchechmedjiev, A. (2012). État de l'art: mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances. JEPTALN-RECITAL 2012, 295.

[**Tur00**] Turney, P. D. (2000). Learning algorithms for keyphrase extraction. Information Retrieval, 2(4), 303-336.

[**UC06**] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna, "Semantic annotation for knowledge management: Requirements and a survey of the state of the art," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 4, no. 1, pp. 14–28, 2006.

[**VM02**] M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna, "MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup," in Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, Sigüenza, Spain, 2002, vol. 2473, pp. 379–391.

[**W3c04**] W3C, "SWRL: A Semantic Web Rule Language Combining OWL and RuleML," 2004. [Online]. Available: http://www.w3.org/Submission/SWRL/.

[**WP94**] WU, Z. et PALMER, M. (1994). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on ACL, volume 2 de ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

[**WY07**] Wan, X., Yang, J., & Xiao, J. (2007, June). Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. InAnnual Meeting-Association for Computational Linguistics (Vol. 45, No. 1, p. 552).

[ZH08] Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. Journal of Computational Information Systems, 4(3), 1169-1180

Annexe

1- Extrait de la classe de nettoyage des documents (java)

```
public class Tokenizer {
   public String Token(String fichier){
     String st2=new String("");
     gu="";
     int lastp=-5;
   int lasto=-5;
   int h=0;
   int ts=0;
   String  st1 = new String("");
    String str=new String("");
     String s=new String("");
     String text;
     String Doc=new String("");
      try {
   BufferedReader r = new BufferedReader(new FileReader(fichier));
   while((st1=r.readLine())!=(null)){ st2=st2+"\n"+st1;    }
    sa =st2;
   for( int i=0;i<st2.length();i++){
    // int e=-1;
   int p= st2.indexOf("</HTML>",i);

    if(p!=-1){st2=st2.substring(0,p);i=st2.length();}
   }
   //System.out.println(st2);
  for( int i=0;i<st2.length();i++){
    int p= st2.indexOf("<!--",i);
   int o=st2.indexOf("-->",i++);
   if((p!=-1)&&(o!=-1)&&(o>=p)){

       text=st2.substring(p+1);
       if(text.equals("<")){}
       else{
          if(p!=lastp &&lasto!=o)
          s=s+st2.substring(h,p);
          h=o+3;
          lastp=p;lasto=o;
       }
     ts=ts+1;}
     if((p==-1)&&(o!=-1)&&(o>=p)){
      text=st2.substring(p+1);
      if(text.equals("<")){}
      else{
        s=s+st2.substring(o+3,st2.length());
         i=st2.length();
      }}
```

```
 }
 if(ts==0){s=st2;}
//System.out.println("   lll          "+s+"                lll");
 for( int i=0;i<s.length();i++){

  int p= s.indexOf(">",i);
  int o=s.indexOf("<",i++);
  if((p!=-1)&&(o!=-1)&&(o>=p)){
     text=s.substring(p+1);
     if(text.equals("<")){}
     else{
        if(p!=lastp &&lasto!=o)
        str=str+s.substring(p+1,o);
        lastp=p;lasto=o;
     }}
  }
  StringTokenizer g =new StringTokenizer(str," , ,\n",false);

   while (g.hasMoreTokens()) {
   String q=g.nextToken();
   gu=gu+" "+q;

 }
 Doc=gu;

 }catch (IOException e) { System.err.println("openFile: " + e);}

  return Doc;
 }
}
```

2- Extrait de la classe de calcul de similarité (java)

```
//package annotation_semantique;
import java.io.*;
import java.util.*;
import fr.inrialpes.exmo.align.impl.method.*;
import net.didion.jwnl.*;
import net.didion.jwnl.data.*;
import net.didion.jwnl.data.list.*;
import net.didion.jwnl.dictionary.Dictionary;
public class Similarity {
   public static final double NOUN_WEIGHT = 0.60;
   public static final double ADJ_WEIGHT = 0.25;
   public static final double VERB_WEIGHT = 0.15;
   // Results tables
   double[][] nounsResults;
   double[][] verbsResults;
   double[][] adjectivesResults;
    Hashtable nouns1 = new Hashtable();
    Hashtable nouns2 = new Hashtable();
    Vector similarity =new Vector();
    Vector similarity2 =new Vector();
    Vector similarity3 =new Vector();
    String s;
   // Weights tables (masks)
   double[][] nounsMasks;
   double[][] verbsMasks;
   double[][] adjectivesMasks;
   /**
    * Initialize the JWNL API. Must be done one time before computing distance
    * Need to configure the file_properties.xml located in the current
    * directory (ontoalign)
    *
    */
   public void Initialize() {
         try {
               JWNL.initialize(new FileInputStream("./file_properties.xml"));
         } catch (Exception ex) {
               ex.printStackTrace();
               System.exit(-1);
         }
   }

public double BasicSynonymDistance(String s1, String s2) {
         double Dist = 0.0;
         double Dists1s2;
         int j, k = 0;
         int synonymNb = 0;
         int besti = 0, bestj = 0;
         int syno = 0;
         double DistTab[];
         IndexWord index = null;
         Synset Syno[] = null;
```

```
        s1 = s1.toUpperCase();
        s2 = s2.toUpperCase();

        Dists1s2 = StringDistances.subStringDistance(s1, s2);

        try {
            // Lookup for first string
            index = Dictionary.getInstance().lookupIndexWord(POS.NOUN, s1);
        } catch (Exception ex) {
            ex.printStackTrace();
            System.exit(-1);
        }
        // if found in the dictionary
        if (index != null) {
            try {
                // get the groups of synonyms for each sense
                Syno = index.getSenses();
            } catch (JWNLException e) {
                e.printStackTrace();
            }
            // number of senses for the word s1
            synonymNb = index.getSenseCount();
            DistTab = new double[synonymNb];
            // for each sense
            for (k = 0; k < synonymNb; k++) {
                // for each synonym of this sense
                for (j = 0; j < Syno[k].getWordsSize(); j++) {
                    Dist = StringDistances.subStringDistance(Syno[k].getWord(j)
                            .getLemma(), s2);
                    if (Dist < Dists1s2) {
                        Dists1s2 = Dist;
                        besti = k;
                        bestj = j;
                    }
                }
            }
        }

        return Dists1s2;
    }
```

3- Extrait de la classe owl (java)

```
  public class OWL{

public void owl(int f){
 String uri1;
 String s=new String();

 onto=new String();
 JenaOWLModel owlModel1 = ProtegeOWL.createJenaOWLModel();
  try {
 Loria u=new Loria();
  s=u.loadObjet2("Patch.SER");
  list3.removeAll();
uri1= "file:/"+s.replace('\\','/');
onto=uri1;
//uri1="file:/D:/Documents@and@Settings/instOntology.owl";
String con=new String();
  String ai=new String() ;

owlModel1 = ProtegeOWL.createJenaOWLModelFromURI(uri1);
String c=new String();
        ai=(String) MOT22.get(f);
        con=ai.substring(0,1).toUpperCase()+ai.substring(1,ai.length());
       c=con;
       if(owlModel1.getInstance(c)==null){
         if(owlModel1.getInstance(c.toLowerCase())==null){
         list3.add("null");}else {
         RDFSClass cls2=owlModel1.getRDFSNamedClass(c.toLowerCase());
          OWLNamedClass Cls2=(OWLNamedClass) cls2;
        for (Iterator it = Cls2.getSuperclasses(false).iterator(); it.hasNext();) {
          RDFSClass superclass = (RDFSClass) it.next();
                    list3.add(superclass.getName());
                          }

       }

       }
      else {
        RDFSClass cls2=owlModel1.getRDFSNamedClass(c);
         OWLNamedClass Cls2=(OWLNamedClass) cls2;
   for (Iterator it = Cls2.getSubclasses(false).iterator(); it.hasNext();) {
        RDFSClass subclass = (RDFSClass) it.next();
                list3.add(subclass.getName());

       }
     }
    }
      catch (Exception e) {
                 e.printStackTrace();
    }
```

Résumé

Le nombre des documents sur le web s'accroit de jour en jour, et la localisation des documents est devenue une lourde tâche surtout lorsqu'il s'agit de recherche d'un contenu. Ajouter une couche sémantique aux documents c'est l'une des méthodes qui donne aux documents plus de sémantique, et alors la recherche devient un sens pas un terme. Donc un document doit être décrit par une liste de concepts reliés par des relations, c'est l'annotation sémantique.

Dans cette thèse, Nous nous intéressons à une approche d'annotation sémantique des documents pédagogiques sur le web. Cette approche vise à annoter un document par contenu et par contexte ; par contenu, le document sera représenté par des mots clés qui seront ensuite reliés à des concepts d'ontologie et par contexte, puisqu'un document cite généralement d'autres documents, nous propageons les annotations des références pour annoter le document citant. Nous appliquerons ensuite un module de validation qui consiste à rendre nos annotations cohérentes.

Mots clés : annotation sémantique, métadonnées, recherche d'information, Ontologie, validation

Abstract

The number of documents on the web is growing day by day, and the location of the documents has become a difficult task especially if it comes to looking for content. Add a semantic layer to words of documents is one of the methods giving more semantics to the documents, and then the research becomes a meaningful, not just words. So a document must be described by a list of concepts linked by relations, it is the semantic annotation. In this thesis, we present a semantic annotation approach of pedagogic documents on the web.

This approach aims to annotate a document by content and context, by content we represents documents by keywords that are connected to the ontology's concepts. By context, as documents cite generally other documents, we propagate the annotations of references to annotate the citing document. We then apply a validation module, which consists to make our annotations consistent

Key words: Semantic annotation, metadata, information retrieval, ontology, validation.

ملخص :

عدد الوثائق على الأنترنت يتزايد يوما بعد يوم و عملية البحث أصبحت مهمة صعبة خاصة إذا ما تعلق البحث بالمحتوى. إضافة طبقة دلالية لمحتوى الوثائق هي واحدة من طرق إعطاءها أكثر دلالة، ومن ثم يصبح البحث بالمعنى لا بالمصطلحات. لذلك يجب أن توصف الوثيقة بقائمة من مفاهيم ترتبط بعلاقات معنوية ،وهذا ما يسمى بالشرح الدلالي.

في هذه الأطروحة، نقدم منهج للشرح الدلالي للوثائق على شبكة الإنترنت. ويهدف هذا النهج إلى إضافة شروحات للوثيقة من حيث المحتوى والسياق. من حيث المحتوى إذ نقوم بتمثيل محتوى المستند بمفردات يتم ربطها بعد ذلك إلى مفاهيم الأنطولوجيا. و من حيث السياق،اعتمادا على مفهوم أن كل وثيقة علمية تحتوي على عدة وثائق مرجعية أخرى، و بالتالي نقوم بتوريث الوثيقة كل شروحات الوثائق التي ذكرتها كمراجع. في آخر خطوة نقوم بتأكيد صحة كل الشروحات الناتجة عن طريق ازالة التكرارات و التناقضات داخل قاعدة الشروحات

الكلمات المفتاحية :الشرح الدلالي ، الفوقية ، البحث عن المعلومات ، أنطلوجيا ، التحقق من الصحة