Introduction :

Ce chapitre présente un aperçu des concepts généraux de notre travail, il débute par quelques détails sur le web sémantique , le W3C , les ressources sémantiques , les ontologies et spécialement DBpédia , ensuite un aperçu sur les langages de ce domaine (RDF , RDFS , SPARQL), en finissant par la définition de la recherche d'information classique et sémantique .

1.1. Le Web Sémantique :

1.1.1. Définition :

Le Web classique se base essentiellement sur la structure du document et les liens entre les pages Web, ce qui rend l'exploitation des informations impossible par les machines. En revanche, le Web sémantique offre une représentation sémantique du contenu, ce qui permet aux agents logiciels l'extraction de l'information.

Dans son ouvrage Weaving the Web, Tim Berners-Lee (co-inventeur avec Robert Cailliau du World Wide Web) donna sa définition la plus aboutie [BER 01] « *Le web sémantique désigne une évolution du web syntaxique. Ce dernier est entièrement destiné à la consultation de documents par des utilisateurs humains. Dans le web sémantique, les contenus sont également compréhensibles et exploitables par des agents logiciels. Sa mise en place permettrait l'automatisation d'un grand nombre de tâches et un accès plus facile à la connaissance, en s'appuyant sur l'utilisation d'un langage de description commun, permettant d'exprimer à la fois des ontologies et des assertions (éventuellement sous formes d'annotations dans des documents web) utilisant le vocabulaire de ces ontologies* ».

Et selon Goble [CHR, KEI 07]: « *le Web sémantique vise a favoriser la découverte, l'échange d'information et l'intégration de l'information par l'étiquetage (tagging) ou le balisage (marking up), et a développer des technologies et infrastructures pour annoter le contenu du Web avec des informations additionnelles appelées " métadonnées " pouvant être traitées par des agents logiciels* ».

D'un point de vue technologique, le web sémantique peut se définir par la pile de standards qui le composent et fixent le cadre de sa mise en place. Un schéma largement diffusé, appelé " semantic web layer cake" (voir Figure 1.1 et Figure 1.2), décrit cette pile de standards et constitue à lui seul un résumé de la feuille de route dressée par Tim Berners-Lee. L'agencement en pile illustre l'organisation et les dépendances entre les standards, chaque couche exploitant les couches inférieures.

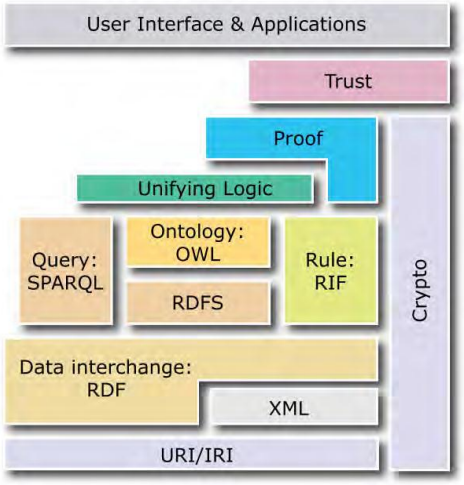


Figure 1.1 : La pile des standards du WS [01].

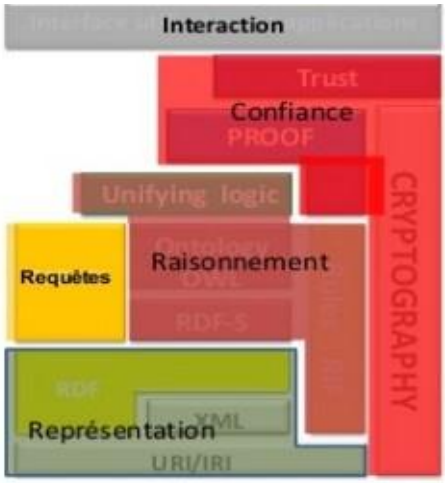


Figure 1.2 : Les groupes de la pile [01]

- URI : permet d'identifier les ressources sur les réseaux.
- XML : est utilisé comme couche de base syntaxique du web sémantique. Le langage XML est considéré comme un standard pour le transport de données sur le web.
- RDF : représente les métadonnées pour les ressources web.
- OWL : fondée sur une formalisation commune, spécifie la sémantique de méta-données fournies dans le web sémantique.
- Logique : s'appuie sur des règles d'inférence qui permettent le raisonnement intelligent exécuté par des agents logiciels.
- Preuve : les langages logiques permettent la mise en œuvre d'outils de raisonnement. Des outils de raisonnement sont disponibles pour des langages comme OWL et permettent par exemple de tester la cohérence des informations, de les classifier.
- Confiance : la couche confiance est située au haut de la pyramide. Elle concerne l'utilisation de signatures numériques et d'autres types de connaissances afin de garantir la fiabilité et l'origine des informations, par des recommandations d'agents de confiance, de la notation, des organismes de certification et des organismes de Consommateurs.

1.1.2. Le World Wide Web Consortium (W3C) :

Abrégé par le sigle W3C, est un organisme de standardisation à but non lucratif, fondé en octobre 1994, chargé de promouvoir la compatibilité des technologies du World Wide Web telles que : (HTML5 , HTML , XHTML , XML, RDF, SPARQL, CSS , XSL,

PNG, XSL, PNG, SVG et SOAP). Fonctionnant comme un consortium international, il regroupe plusieurs entreprises partenaires.

Protocoles utilisés :

a- le protocol HTTP :

Le HTTP (*HyperText Transfer Protocol*), permet, à partir d'une adresse URL, de demander une représentation de la ressource identifiée (par exemple une page Web) et localisée par cet URL, et d'obtenir en retour soit les données de cette représentation, soit des codes d'erreur indiquant un problème rencontré. Nous avons tous déjà rencontré la célèbre « erreur 404 » qui indique que la page demandée n'a pas été trouvée.

b- les U R I :

Un URI (*Uniform Resource Identifier)* est un identifiant unique sur le Web (*une* adresse page web à propos de la ressource). Cet identifiant est le plus souvent sous la forme d'une URL de page web. Un URI est dit *" déréférençable*s " il désigne une page web existante à laquelle un utilisateur peut accéder via un navigateur Web.

c- le langage XML :

Le XML *(eXtensible Markup Language*), traduit ainsi (Langage *à balises étendu*), est un langage HTML amélioré permettant de définir de nouvelles balises. Il permet de mettre en forme des documents grâce à des balises (markup). XML peut être considéré comme un métalangage permettant de définir d'autres langages, c'est-à-dire définir de nouvelles balises permettant de décrire la présentation d'un texte. La force de XML est dans sa capacité de décrire n'importe quel domaine de données grâce à son extensibilité. Il va permettre de structurer, poser le vocabulaire et la syntaxe des données qu'il va contenir. Les balises XML décrivent le contenu plutôt que la présentation.

Le XML permet de séparer le contenu de la présentation .ce qui permet par exemple d'afficher un même document sur des applications ou des périphériques différents sans pour autant nécessiter de créer autant de versions du document que l'on nécessite de représentations .

1.1.3 Annotation sémantique et concept de métadonnée :

a . Annotation sémantique :

Le terme « *Annotation* » est défini dans le dictionnaire comme étant « *une note explicative ou critique qui accompagne un texte- une note de lecture que l'on inscrit sur un livre*

». Ce terme dérive du terme latin « *Annotare* » qui signifie « *accompagner un texte de notes, de remarques et commentaires* », et parce que nous annotons un sujet, une annotation seule n'a aucun sens, elle est associée a l'objet qui a été annoté.

Nous nous intéressons principalement à l'annotation de document par le contenu, l'un des aspects du web sémantique est de pouvoir créer et manipuler les annotations sémantiques, ce sont des Métadonnées de documents (voir l'exemple du Figure 1.3)

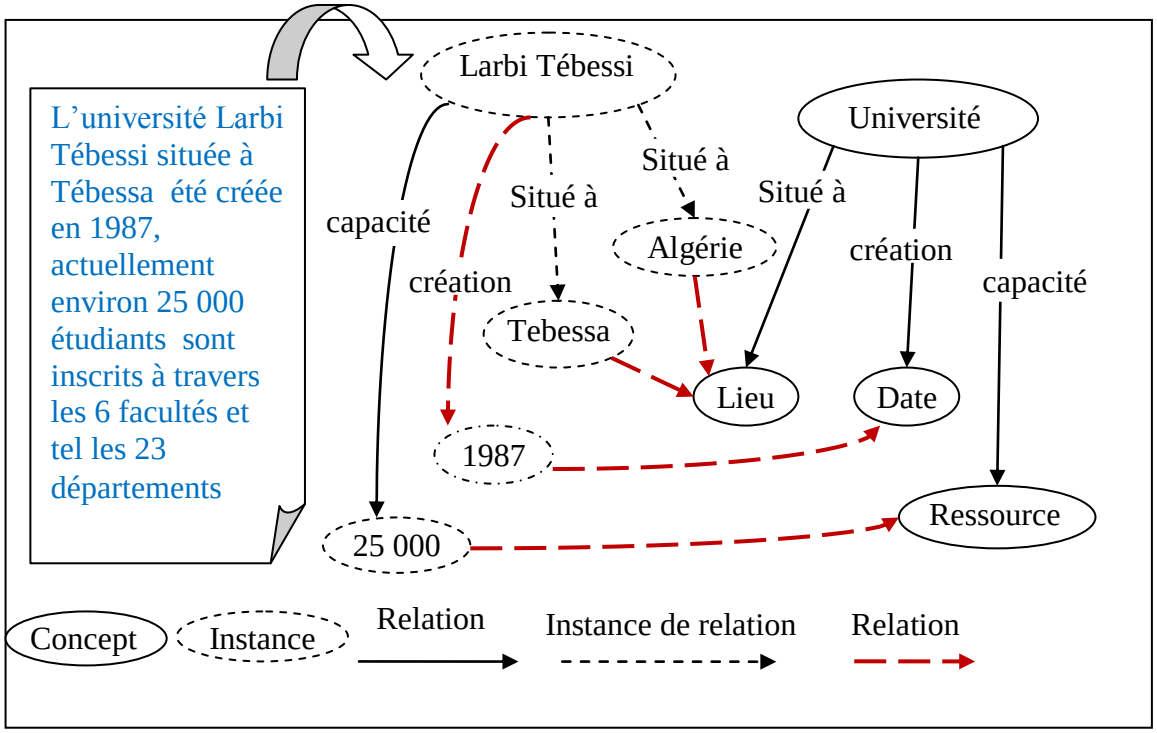


Figure 1.3 : Exemple d'Annotation Sémantique basée ontologie

La sémantique de l'annotation est fondée sur des vocabulaires dans les ontologies spécifies explicitement par un langage de représentation, par conséquent l'annotation sémantique est le processus le plus adapté pour le partage du contenu des documents.

Une annotation sémantique peut alors être définie selon Tuan Dung **[**TUA 06] comme : « *Une représentation formelle d'un contenu, exprimée par des concepts, des relations, et instances décrits dans une ontologie reliée à le ressource documentaire* ».

Les approches d'annotation sémantique se distinguent par les types de connaissances ontologiques utilisés, et par la nature et la granularité des parties de document annotées. Les éléments ontologiques utilisés pour l'annotation peuvent être classés en cinq classes:

- Les concepts : termes définis dans une ontologie.

- Les instances de concepts.
- Les relations.
- Les instances de relations.
- Les valeurs littérales.

b. Concept de métadonnée :

Le Web sémantique est un moyen d'échange et de partage des ressources grâce à la représentation sémantique du contenu. Ces ressources sont décrites par des informations structurées additionnelles " *les métadonnées* ".

Les métadonnées selon Jorge Cardoso et Amit Sheth [JOR, ADM 06] sont : « *des données sur les données* ».

Le préfixe « *méta* » signifie que ces informations ne concernent pas le sujet traité dans le document mais qu'il s'agit de données portant sur le document lui même. Ces métadonnées peuvent être, suivant les technologies utilisées, intégrées dans le document, ou bien rassemblées en dehors des documents, dans une base de métadonnées jouant le rôle d'un catalogue de bibliothèque.

L'utilisation des métadonnées permet de décrire, identifier et définir le contenu de la ressource afin de rendre son contenu exploitable. Les métadonnées peuvent exister sur différents niveaux distincts. Les informations décrivant la syntaxe, la structure et le contexte sémantique.

Avec la diffusion de l'internet, et la croissance du Web, le terme " *métadonnées* " commence à être utilisé dans le contexte de la description des ressources numériques disponibles. Ainsi que les documents textuels qui sont compréhensibles par les humains, ils ont besoin des métadonnées pour que les machines puissent les manipuler, les contrôler, les rechercher et les récupérer.

La figure 1.4 illustre un exemple des métadonnées pour un fichier PDF.

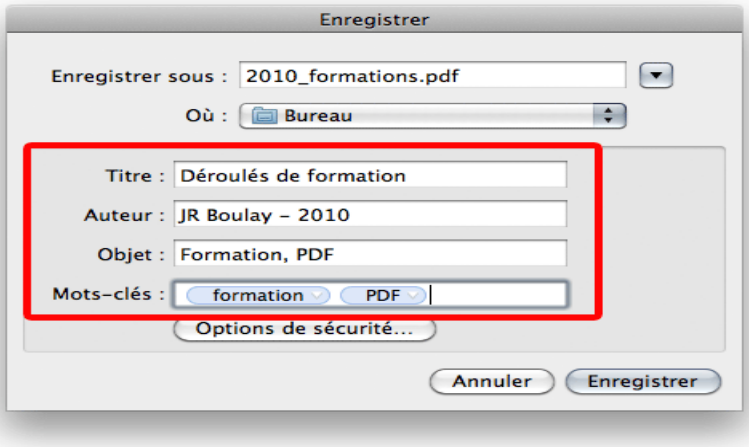


Figure 1.4 : Exemple de Métadonnées d'un fichier PDF. [02]

1.1.4 . Les Ressources Sémantiques (" Les Ontologies "):

Une ressource sémantique est un vocabulaire contrôlé. C'est une liste de termes d'un domaine ou de plusieurs domaines qui ont été énumérés explicitement. Cette liste est contrôlée par un expert. Chaque terme du vocabulaire doit avoir (théoriquement) une définition non ambigüe et non redondante (mais dans la pratique n'est pas toujours vrai).

Ils existent différents types de ressources sémantiques telles que les taxonomies, les thesaurus, les ontologies, glossaires et dictionnaires. Les gens utilisent dans la plupart du temps le mot ontologie pour indiquer ces différents concepts.

a. Taxonomie :

C'est la technique utilisée pour faire un classement dans un système hiérarchique des organismes. Une taxonomie est une collection de termes organisés en une structure hiérarchique. Les termes dans une taxonomie sont reliés entre eux par des relations parent-enfant.

Il peut y avoir différents types de relations parent-enfant dans une taxonomie (par exemple, tout-partie, genre-espèces, Type-Instance), mais une bonne pratique limite les relations parent-enfant (voir Figure 1.5).

b. Thesaurus :

Un thésaurus est un réseau de termes d'un vocabulaire contrôlé. Il utilise d'autres relations entre les termes, en plus de la relation parent-enfant.

Dans le domaine de recherche d'information un thésaurus cherche à dicter les relations sémantiques entre les termes pendant la phase d'indexation. Un thésaurus sert à minimiser l'ambigüité sémantique en garantissant l'homogénéité et la cohérence dans le stockage et la récupération des termes du contenu.

Un thésaurus sert à guider à la fois la phase d'indexation et la phase de recherche en sélectionnant le meilleur terme ou la meilleure combinaison de termes qui représente un sujet donné. (Voir Figure 1.5)

c. Ontologie :

Face à l'émergence des sources d'informations disponibles de plus en plus nombreuses et complexes, il est nécessaire de permettre une description de ces informations non seulement en termes de structure (aspect syntaxique) mais également en termes de signification (aspect sémantique). La description de métadonnées sur les sources d'informations prenant en compte la description structurelle mais aussi sémantique des informations est un problème important. Cette sémantique peut être exprimée à l'aide d'ontologies.

Dès les début des années 90, Gruber proposait la définition suivante : " *Une ontologie est une représentation formelle des connaissances* " [GRU, 93]. Ensuite, il complète en précisant qu'une ontologie est une "*spécification formelle et explicite d'une conceptualisation partagée*" [GRU, OLS 94].

Et selon Gomez Carpio [GOM 10], " *Une ontologie permet la modélisation d'un domaine de connaissances et peut être vue comme un modèle conceptuel d'un domaine particulier, qui décrit les concepts de ce domaine et les relations entre ces concepts. Elle est généralement considérée comme une base de connaissances et est au centre des développements émergents du Web sémantique* ".

Les ontologies permettent la modélisation d'informations agréées par une communauté de personnes et accessibles par une machine pour développer des services automatisés et par conséquent, jouent un rôle de référence pour décrire la sémantique des informations à partager.

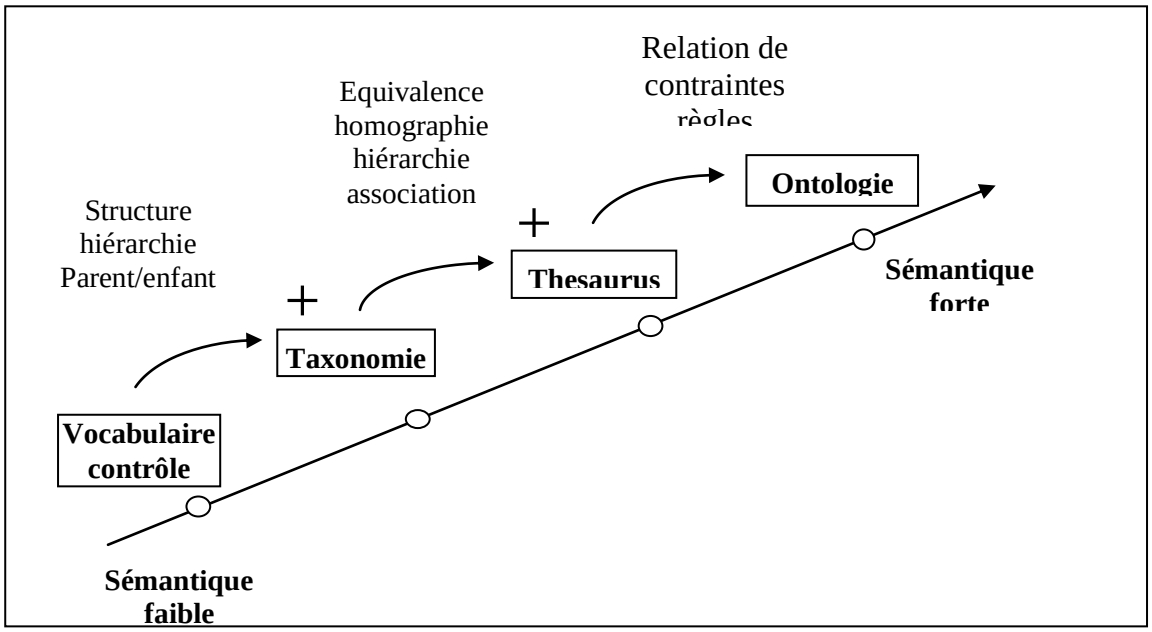


Figure 1.5 : Vocabulaires et richesse sémantique.

- **Types d'ontologies :**

Selon Van Heijst [VAN 97], les typologies d'ontologies fondées sur le sujet de la conceptualisation comportent quatre catégories :

- ✓ Les ontologies d'application : elles contiennent toutes les informations nécessaires pour modéliser les connaissances pour une application particulière.
- ✓ Les ontologies de domaine : elles fournissent un ensemble de concepts et de relations décrivant les connaissances d'un domaine spécifique.
- ✓ Les ontologies génériques (dites aussi de haut niveau) : elles sont similaires aux ontologies de domaine, mais les concepts qui y sont définis sont plus génériques et décrivent des connaissances tels que l'état, l'action, l'espace et les composants. Généralement, les concepts d'une ontologie de domaine sont des spécialisations des concepts d'une ontologie de haut niveau.
- ✓ Les ontologies de représentation (méta-ontologies) : elles fournissent des primitives de formalisation pour la représentation des connaissances. Elles sont généralement utilisées pour écrire les ontologies de domaine et les ontologies de haut niveau.

- **L'utilité des Ontologies :**

Les ontologies selon Domenico M. Pisanelli et ces co-équipiers [DOM, ALD, GER 02] servent à :

- ✓ Créer une compréhension partagée pour unifier différents points de vue,
- ✓ Faciliter la communication entre les personnes impliquées dans la construction du SI,
- ✓ Permettre la réutilisation des connaissances du domaine, faciliter la récupération, l'intégration et l'interopérabilité entre les sources hétérogènes de la connaissance,
- ✓ Fournir une base de représentation des connaissances du domaine,
- ✓ Aider à identifier les catégories sémantiques de domaine.

- **Ontologies Majeures :**

Les Communautés actives autour du web sémantique sont reconnues par *" Les Bases Majeures "* , certaines de ces bases sont particulièrement connues et utilisées, elles sont liées à un très grand nombre d'autres bases, à un tel point qu'on peut les qualifier de « *majeures* » dans le W3C : (*DBpedia , YAGO , FOAF , Freebase ….. etc*) (voir Figure 1.6).

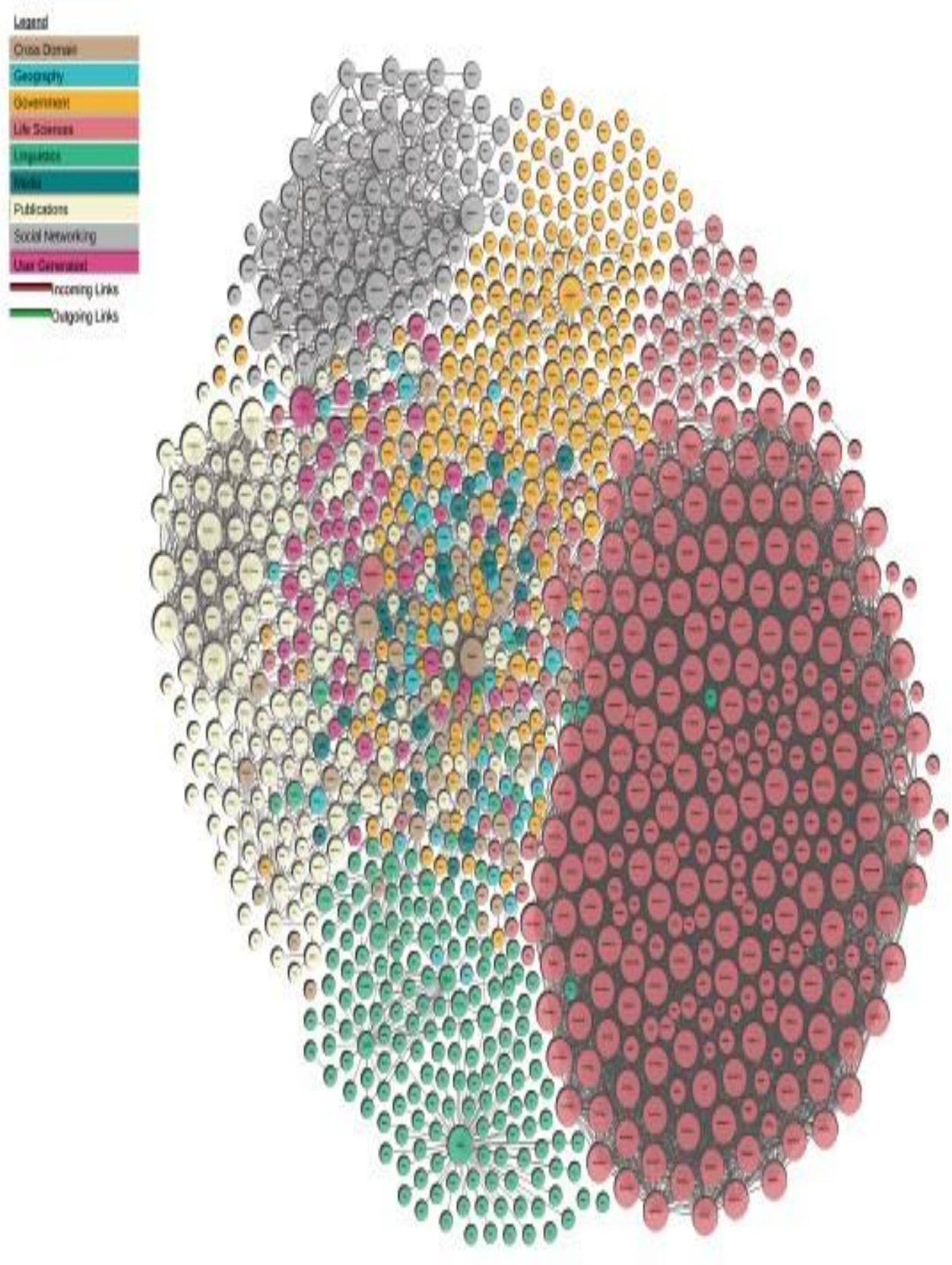


Figure 1.6 : Le web de données en Avril 2017. [03]

- **Web de données liées (Linked Data) :**

L'objectif de Linked Data est d'utiliser le Web pour faciliter la création de liens entre des données connexes qui n'étaient pas encore liées, ou qui l'étaient déjà via d'autres méthodes.

- **Web de données ouvertes et liées (Linking Open Data)** :

Linking Open Data est un projet communautaire collectif visant a publier un maxi--mum de données libres de droits sur le Web de données selon les principes de Linked Data, en interconnectant des données de différentes sources de données. Une fois ce réseau de sources de données créé, il est alors intéressant de développer des clients Linked Data permettant de naviguer parmi ces données interconnectées.

1.1.5 . L'ontologie DBpedia :

Depuis Juin 2007, une nouvelle application est en train de révolutionner la manière dont en peut rechercher et utiliser des données. DBpedia [LET, JEN, ISE 15] (voir Figure 1.7), est en effet une version base de données libre (open source) de certaines informations de Wikipédia avec une structure intelligible et partagée. Les données sont principalement extraites des boîtes d'informations (dites *infobox*) présentes dans les articles de Wikipédia.

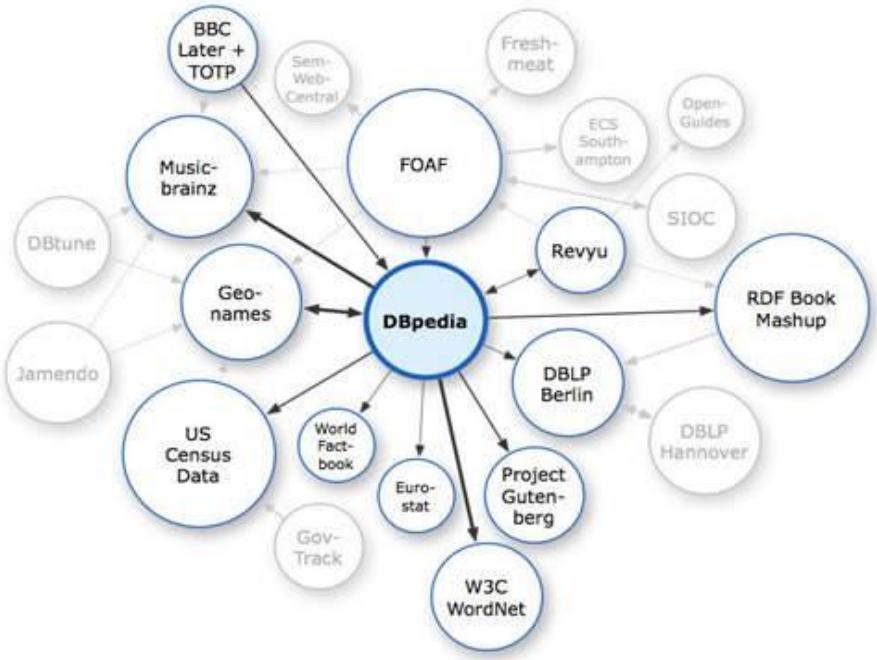


Figure 1.7 : L'Ontologie DBpedia . [04]

Le contenu de DBpedia [LET, JEN, ISE 15] est obtenu du moissonnage des pages et des mises à jours de Wikipedia. Le système de moissonnage à l'origine de DBpedia crée des triplets principalement à partir des cadres de résumés formatés, les « infoboxes», des pages de Wikipedia et en moindre proportion, du traitement du langage naturel des textes d'articles.



Figure 1.8 : Wikipedia avec infobox pour créer DBpedia. [05]

La base DBpedia est mise à jour annuellement (voir Tableau 1.1). Cette action est faite depuis la base DBpedia Live, miroir de DBpedia, qui est mise à jour directement à partir de Wikipedia.

La première version de DBpedia [LET, JEN, ISE 15] était extraite depuis la version anglophone de Wikipédia, les extracteurs d'informations n'étant conçus que pour être appliqués sur les infoboxes de cette version. Ce choix a eu pour conséquence que seules les entrées de la version anglophones bénéficiaient d'une représentation sous forme de données dans DBpedia .

Mais la création de données d'après une encyclopédie anglophone a rapidement montré des limites de couverture. Il a donc été décidé de poursuivre le développement d'éditions linguistiques de DBpedia.

En mars 2012, Une version entièrement francophone, agrémentée de fonctionnalités nouvelles, a été officiellement révélée. Elle est développée par l'Inria avec le soutien du Ministère de la Culture et l'expertise de l'association Wikimédia France.

Le but principal de cette base de connaissances est de combler ce manque de données gratuites dans la toile, aussi de développer les technologies (web sémantique) pour effectuer des recherches dans ces bases et/ou faire interagir et partager les données récupérées. DBpedia nous permet aussi de faire des recherches avancées sur la structure de Wikipedia. [06]

Années	Nombre Concepts (éléments)	Dont Personnes	Dont Places	Dont Organisations
2010	3.40 Millions	312 000	413 000	140 000
2012	3.70 Millions	416.000	526.000	190 .000
2014	4.58 Millions	1.450.000	735.000	241.000
2016	4.60 Millions	1.500.000	810.000	275.000

Tableau 1.1 : L'évolution de version Anglaise de DBpedia [09]

1.2 . Les Langages des Ontologies :

1.2.1 – RDF :

RDF (Resource *Description Framework)* est à la base des technologies du Web des données. Il s'agit d'un modèle de données simple pour la représentation de connaissances sur le Web dans des *documents RDF*.

Comme son nom l'indique, RDF sert à décrire des *ressources RDF*. Une ressource RDF est un concept ou un élément concret (toute sorte de chose, une personne, un pays, un évènement, ...). Chaque ressource RDF est – théoriquement – unique dans tout le Web des données. Les descriptions des ressources RDF sont composées de *triplets RDF*, (sujet, relation, objet), qui peuvent être représentés sous forme de graphe. Un triplet RDF (voir Figure 1.9) exprime une relation entre un sujet (la ressource décrite) et un objet. Autrement dit, un triplet décrit une propriété du sujet ayant pour valeur l'objet du triplet. La description d'une ressource correspond à l'ensemble des triplets qui contiennent la ressource en sujet.

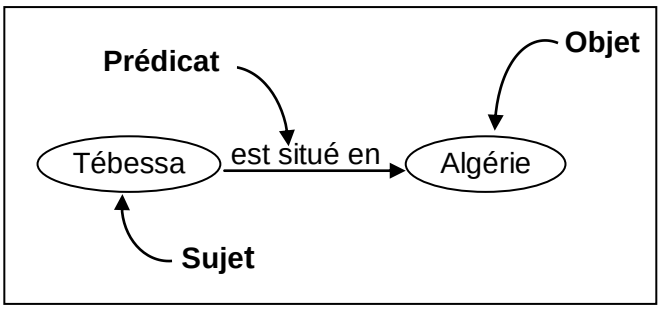


Figure 1.9 : Représentation d'un triplet RDF (sujet-relation-objet).

Sujet : est une ressource, soit identifiée par un URI, sinon correspondant à un nœud blanc.

Relation : est toujours une ressource identifiée (*URI*), aussi appelée "*prédicat*".

Objet : soit une ressource (identifiée ou non), soit une donnée brute, aussi appelée *littéral*.

Un document *RDF*, (voir Figure 1.10 et Figure 1.11), est également appelé " *ensemble de triplets RDF* " ou encore " *graphe RDF* " . Il peut en effet être représenté sous forme d'un multi-graphe orienté étiqueté.

< Ahmed><parent de ><Omar>
< Ahmed><parent de ><Lina >
< Ahmed><habite ><Algérie>

Figure 1.10 : Ensemble de triplets RDF.

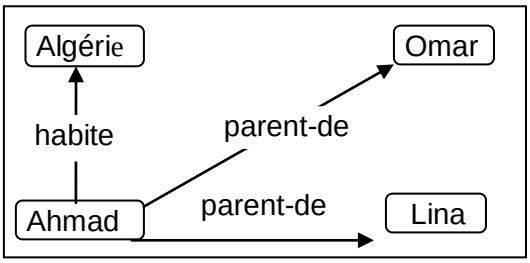


Figure 1.11 : Représentation graphique des triplets RDF du Figure 1.10

- **Les Syntaxes de RDF :**

Un document RDF peut être stocké dans des fichiers en suivant une des nombreuses syntaxes de RDF. Ces syntaxes permettent de représenter des documents RDF de différentes manières sans en changer le contenu de façons plus ou moins adaptées pour l'exploitation par la machine ou par l'humain.

La syntaxe RDF/XML est la première syntaxe à être apparue pour l'écriture de documents RDF. Elle a été introduite en même temps que RDF, basée sur XML et se place dans la continuité des technologies du Web. Son caractère très verbeux, hérité de XML, la rend peu adaptée à l'exploitation par l'humain. Les fichiers RDF/XML, dont un exemple est en Figure 1.12, utilisent les extensions ".rdf ".

```
<rdf:RDF>
<rdf:Descriptionrdf:about="http://www.exemple.com/rdf#Lina">
<ex:parentDerdf:resource="http://www.exemple.com/rdf#Ahmad"/>
<ex:parentDerdf:resource="http://www.exemple.com/rdf#Omar">
<ex:habiterdf:resource="http://www.exemple.com/rdf#Algerie"/>
</rdf:Description>
</rdf:RDF>
```

Figure 1.12 : Document représenté en Figure 1.10 au format RDF/XML.

- **Les Graphes Conceptuels (GCs) :**

Selon GOMEZ Carpio [GOM 10] , Le modèles des graphes conceptuels « G.C » introduits par Sowa en 1984 : « *est un modèle formel et expressif de représentation des connaissances d'un domaine d'application* ».

Les G.C est un langage intermédiaire entre des formalismes destinés aux traitements automatiques et le langage naturel.

Ils tiennent leurs origines des réseaux sémantiques de Quillian 1968 [GOM 10], qui depuis plusieurs extensions ont été proposées, aussi un graphe conceptuel n'a pas de sens propre, c'est au travers des réseaux sémantiques que les concepts et les relations sont rattachés au contexte. Le graphe conceptuel illustré par la Figure 1.13, correspond à la phrase « *Chaque employé est recruté par quelques managers à certaines dates* ».

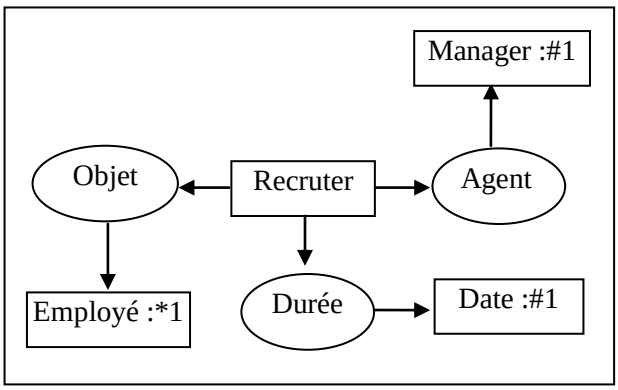


Figure 1.13 : Exemple de graphe conceptuel.

1.2.2- RDF(S) (RDF Schema) :

Les ontologies sont des ensembles de concepts permettant de définir des regroupements de ressources, les classes, de définir les relations utilisables pour les décrire, les propriétés, et de les hiérarchiser.

Le W3C a fourni en 2004 un premier langage basique RDFS (RDF *Schema*) , permettant de définir des ontologies simples. RDFS a ensuite été légèrement étendu en 2014 dans sa version 1.1, en même temps que RDF 1.1. Traditionnellement, on sépare dans un document RDF :

- ✓ les définitions des classes et des propriétés, qui forment l'ontologie,
- ✓ les définitions et mises en relations des instances des classes (aussi appelés individus), qui forment les faits du document.

Les bases RDF du Web des données peuvent utiliser plusieurs ontologies pour définir leurs données, cependant lorsque nous discutons des données d'une seule base, nous considérons l'union des ontologies utilisées dans la base comme formant « l'ontologie de la base ».

- ➢ **Les Classes :**

Une classe est un groupe de ressources ayant des propriétés similaires, Personne, Voiture, Pays, …… *etc.* On dit d'une ressource appartenant à une classe qu'elle est *instance* de la classe. Une instance d'une classe est une ressource liée par la relation rdf:type à une classe. Les classes définies en RDFS sont elles-mêmes des instances de rdfs:Class.

Les classes peuvent être organisées en hiérarchies pour indiquer des classes plus générales et des sous-classes plus spécialisées. Dans ces hiérarchies, on considère que les instances des sous-classes sont également instances des classes plus générales. Les hiérarchies de classes sont définies par la relation rdfs:subClassOf, tel qu'en Figure 1.14.

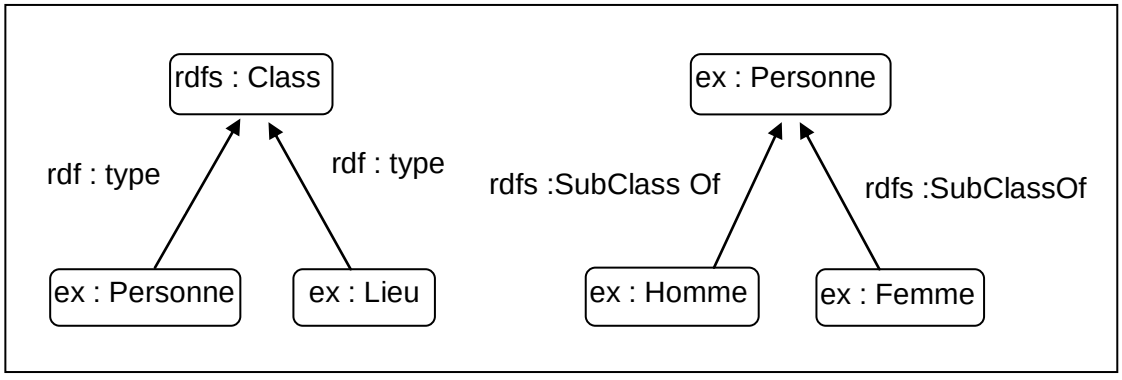


Figure 1.14 : Exemple de l'hiérarchie de classes.

Le langage RDFS définit des ressources utiles à la définition de classes, notamment :

- ✓ rdfs:Resource est la classe par défaut de toutes les ressources définies selon RDFS.
- ✓ tous les types de données des littéraux sont par défaut instances de la classe rdfs:- Datatype.
- ✓ toutes les relations sont instances de la classe rdf:Property. Nous considérons de façon équivalente les termes « relation » et « propriété ».

➢ Les propriétés :

Les propriétés sont les relations de triplet qui sont définies dans l'ontologie. Toute propriété d'une ontologie définie par le langage RDFS est instance de la classe rdf:Property. Tout comme les classes, les propriétés peuvent être organisées en hiérarchies par la relation rdfs:subPropertyOf, voir par exemple Figure 1.15. Une hiérarchie de propriétés définit des niveaux de généralisation ou de spécialisation entre relations. On considère que deux ressources liées par une propriété sont également liées par ses propriétés plus générales.

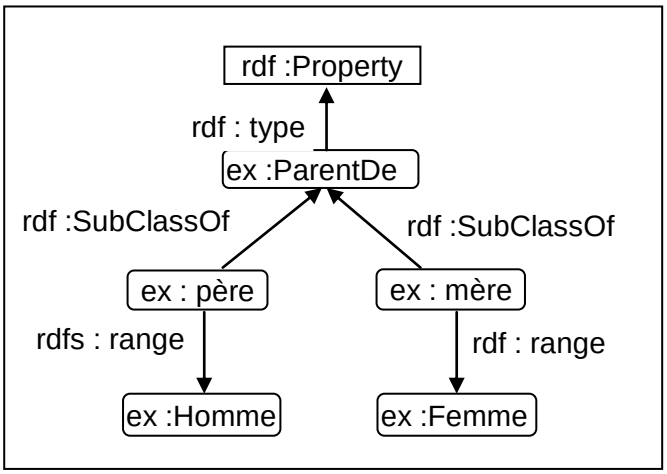


Figure 1.15 : Exemple de hiérarchie de propriétés

➢ Les Individus :

Un individu est une ressource instance d'une classe définie dans l'ontologie – en dehors des classes définies par RDFS – dont des exemples sont donnés en Figure 1.16. Pour un individu, la relation d'instanciation peut être traduite en langage naturel par « l'individu est un ... » ou « l'individu est une sorte ». Par définition, un individu est sujet d'au moins une relation d'instanciation, il est donc toujours représenté par un URI ou un nœud blanc, jamais par un littéral. Un individu instance d'une classe A est également appelé « individu de type A ».

Les individus sont les éléments de plus bas niveau d'un document RDF. En effet, les langages du W3C ne permettent pas de définir des relations de hiérarchisation entre individus. Selon le langage RDFS, toutes les ressources d'un document sont instances de rdfs:Resource. Ainsi, tous les individus d'un document ont toujours un type en commun.

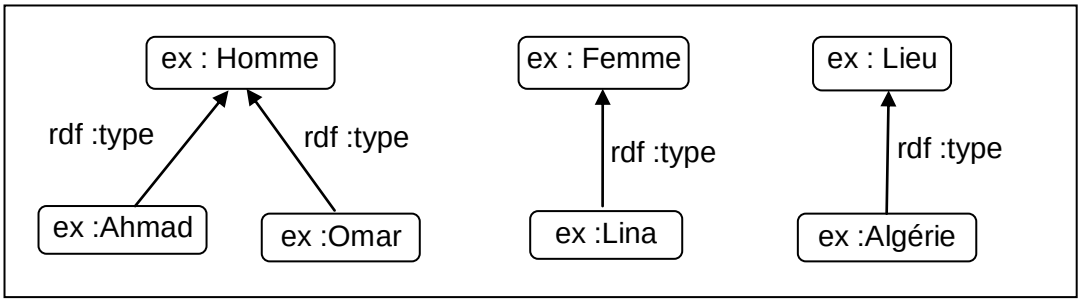


Figure 1.16 : Exemple d'instances définis en figures 1.14 et 1.15.

1.2.3- OWL :

Le langage OWL est apparu en 2004, en même temps que RDF et RDFS, il a ensuite été étendu en 2009-2010 dans la norme OWL 2 .

OWL fait étendre RDFS, il introduit dans les documents RDF une sémantique logique tirée des logiques de description. OWL apporte un langage plus riche pour la description de ressources en permettant la définition de relations entre les classes et entre les propriétés, comme l'équivalence, la différence, l'union, le complément, *etc*...., ainsi que des raisonnements suivants les logiques de description.

1.2.4 - SPARQL :

L'interrogation d'une base RDF se fait très généralement par le langage SPARQL – *SPARQL Protocol and RDF Query Language* – défini par le W3C pour la première fois en 2008 (4 ans après RDF), puis étendu en 2013. Ce langage est un langage d'interrogation de bases de connaissances – à comparer avec le langage d'interrogation des bases de données relationnelles SQL – adapté pour la structure sous forme de triplets des bases RDF.

Une requête SPARQL possède un entête et, le plus souvent, un corps de requête. L'entête contient le mot-clé de la requête – SELECT, ASK, INSERT, *etc.* – avec les éléments demandés en résultat de la requête. Le corps de requête contient un ensemble de « *triplets de recherche* ». Dans les triplets de recherche, un élément sujet, relation ou objet peut être remplacé par une variable, sous la forme d'un nom de variable préfixé du caractère ?. Un ensemble de triplets de recherche forme un *motif de graphe*. On appelle solution d'un motif de graphe un ensemble de données RDF qui forme un morphisme du motif. Ce morphisme substitue les variables du motif par des ressources de façon à ce que le motif de graphe et l'ensemble de données RDF soient identiques. Le corps d'une requête commence par le mot clé WHERE suivi d'un motif de graphe entre accolades ou de plusieurs motifs organisés par des opérateurs.

La syntaxe d'une requête d'interrogation SPARQL est comparable à celle du langage SQL pour les bases de données relationnelles.

Les requêtes d'interrogation SPARQL sont de 4 sortes différentes :

a- **Une requête SELECT** : extrait des résultats sous forme d'un tableau, comme avec SQL, contenant les ressources solutions à un motif de graphe donné et correspondantes aux variables présentes dans l'entête , tel que dans les Figures 1.17 et 1.18 si dessous .

```
SELECT ?adresse
WHERE {
ex:Jeanex:habite ?adresse .
        }
```

Figure 1.17 : Exemple de requête SPARQL SELECT.

```
---------------
| adresse   |
=========
| ex:Paris  |
```

Figure 1.18 : Résultat de la requête SPARQL SELECT

b- **Une requête ASK** retourne une valeur booléenne s'il existe au moins une solution à un motif de graphe donné, tel qu'en exemple de la Figure 1.19 (dont le résultat est " true ").

```
ASK
WHERE
{
ex:Paulex:parentDe ?enfant .
}
```

Figure 1.19 : Exemple de requête SPARQL ASK.

c- **Une requête CONSTRUCT** extrait les solutions à un motif de graphe donné sous la forme d'un nouveau graphe RDF, tel que dans les exemples 1.20 et 1.21. Dans cette requête, les variables sont utilisées dans un motif de graphe en entête pour générer de nouveaux triplets RDF à partir des résultats.

```
CONSTRUCT {
 ex:Julieex:parentDe ?enfant .
 } WHERE {
 ex:Paulex:parentDe ?enfant .
 }
```

Figure 1.20 : Exemple de requête SPARQL CONSTRUCT.

```
ex:Julieex:parentDeex:Jean .

ex:Julieex:parentDeex:Marie .
```

Figure 1.21 : Résultat de requête SPARQL CONSTRUCT

d- **Une requête DESCRIBE** ce type de requête (voir Figure 1.22) retourne des informations sur les résultats de la requête. Les informations décrivant les résultats, qui vérifient les contraintes spécifiées dans la requête, sont retournées sous forme de graphe (voir Figure 1.23). Ce graphe est déterminé par le processus d'interrogation de SPARQL. L'utilisateur ne peut déterminer à l'avance la structure du graphe résultat ;

```
DESCRIBE ex:Paul
```

Figure 1.22 : Exemple de requête SPARQL DESCRIBE

```
SPARQL DESCRIBE.
ex:Paulex:habiteex:Paris .
ex:Paulex:parentDeex:Jean .
ex:Paulex:parentDeex:Marie .
```

Figure 1.23 : Résultat de requête SPARQL DESCRIBE

1.3. La Recherche d'Information :

1.3.1. Introduction :

Le domaine de recherche d'information a pris un grand essor avec l'apparition et le développement de l'internet, où une quantité exponentielle de documents de tout genre (texte, page web, image, vidéo ….) .

Les documents textuels peuvent êtres sans structuration (un texte plat) ou avec une structure tels que les documents XML. ce qui a aidé les systèmes de recherche à améliorer leurs résultats en prenant en compte cette structure.

Un axe récent dans le domaine de recherche d'information qui est la recherche sémantique a pris une place importante ces dernières années (voir Figure 1.24).

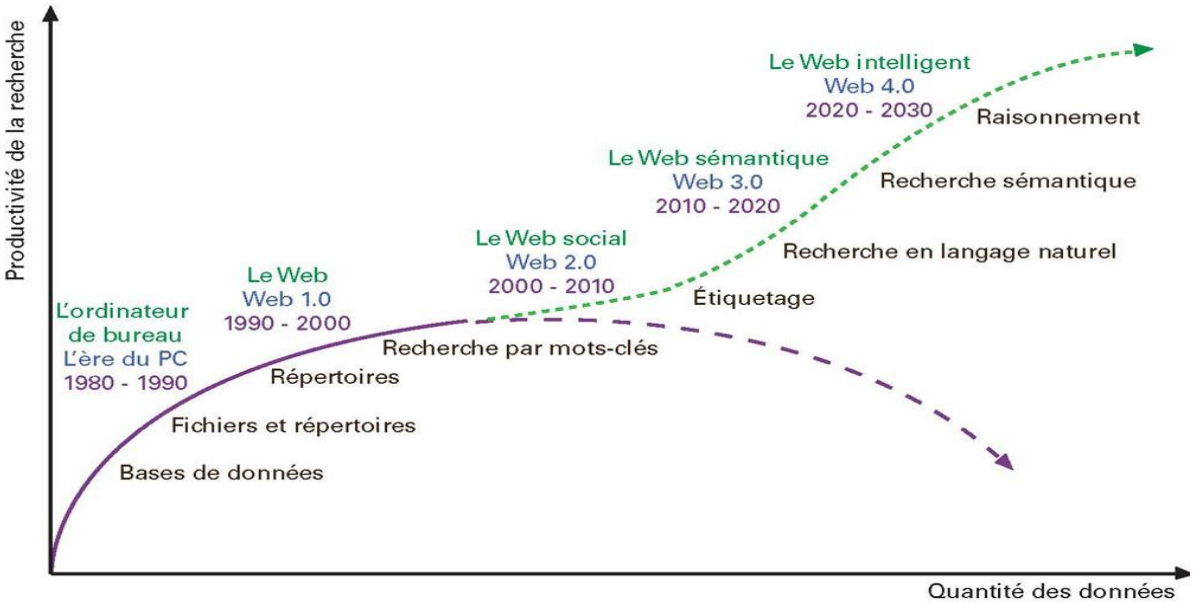


Figure 1.24 : Productivité de la recherche [07]

1.3.2. Définition :

La recherche d'information (RI) est définit par Christopher D. Manning et Prabhakar Raghavan [CHR,MAN,PRA 09] comme: " *c'est l'opération de trouver un matériels (habituellement un document) de nature non structuré (habituellement un texte) qui satisfait un besoin d'information à partir d'une grande collection (généralement stockée sur les ordinateurs)* ".

L'objectif des systèmes de recherche d'information est de retourner des informations (sous forme de document ou une partie de document) selon le besoin de l'utilisateur. Ce besoin est formulé sous forme d'une requête.

Le domaine de (RI) traite les concepts suivants : la représentation, le stockage et l'accès aux documents. La recherche d'information est devenue ces dernières années le moyen le plus utilisé pour l'accès à l'information, principalement en raison de l'augmentation considérable de documents dans le web.

Les systèmes de recherche d'information classique (voir Figure 1.25) sont composés de trois fonctions principales: l'indexation et la représentation des documents, la représentation de la requête utilisateurs, et l'appariement requête-document.

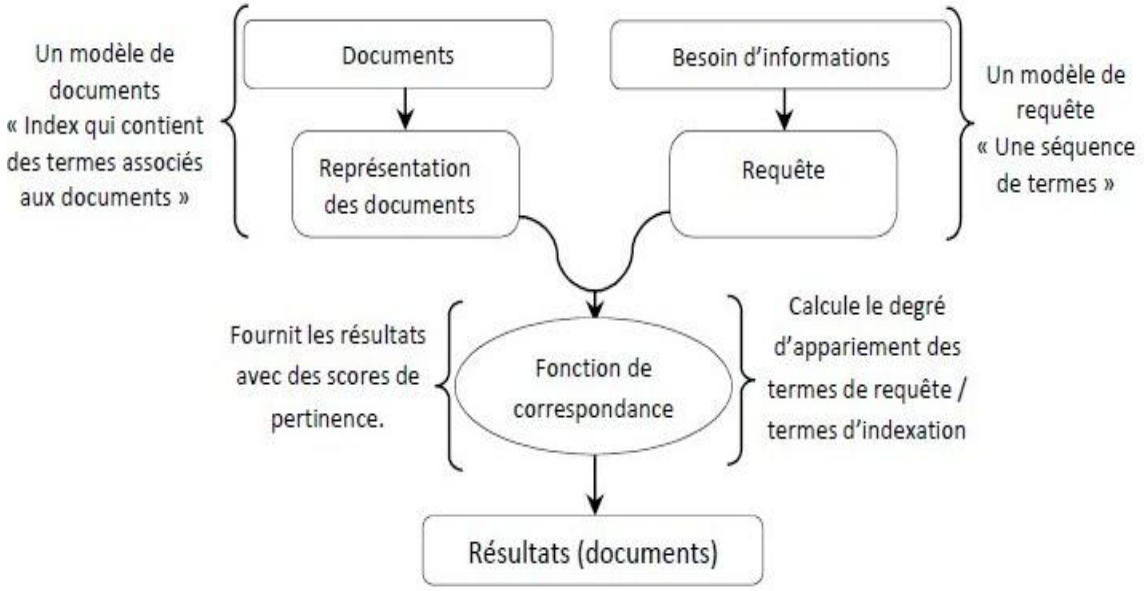


Figure 1.25 : Système de recherche d'information [BOU 11]

1.3.3. Modèles de recherche d'informations :

Un modèle de recherche d'information spécifie pour le système qui l'utilisera les démarches à suivre pour l'accomplissement des étapes décrites précédemment. Ce sont la représentation des documents et des requêtes, l'indexation, la recherche proprement dite, l'évaluation des correspondances documents-requêtes et si besoin en est, la reformulation de la requête.

a. Le modèle booléen :

Le modèle booléen est le modèle classique le plus ancien pour la recherche d'information. Ce modèle est basé sur l'utilisation de la logique booléenne traditionnelle et la théorie des ensembles pour identifier les documents qui correspondent à une requête.

Les documents et les requêtes sont représentés sous formes des expression logique, ou les connecteurs logique sont utilisés (ET, OU, NON). Ce modèle permet de faire un appariement exact entre la requête et les documents en déterminant si les termes de la requête existent ou non (voir Figure 1.26).

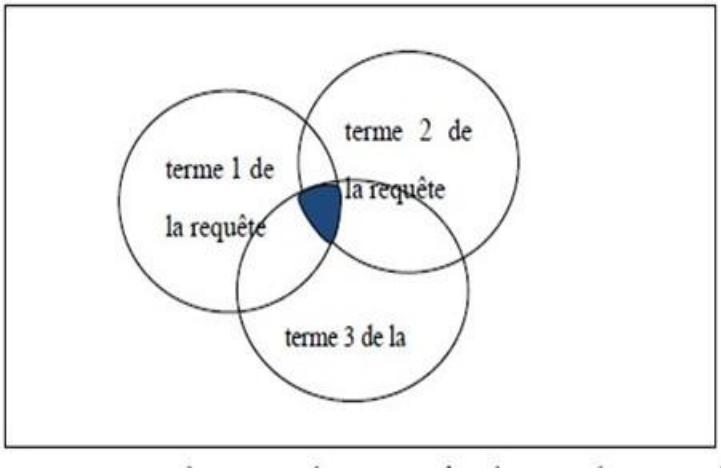


Figure 1.26 : Appartenance des termes d'une requête dans un document.

Ce modèle est simple et très facile à réaliser, mais son inconvénient majeur c'est qu'il ne retourne pas des documents pertinents qui correspondent partiellement à la requête, comme il ne permet pas de classer les documents selon leur degré de pertinence.

b. Le modèle vectoriel :

Le modèle vectoriel (voir Figure 1.27) est l'un des plus utilisé dans la recherche d'information. Dans ce modèle, les documents et les requêtes sont considérés comme des vecteurs dans un espace de terme.

Pour calculer la similarité entre le vecteur requête et le vecteur document, on utilise la fonction de distance euclidienne (similarité cosinus). Après le calcule de la distance entre chaque document et la requête, un document sera classé au premier rang s'il est plus proche du vecteur

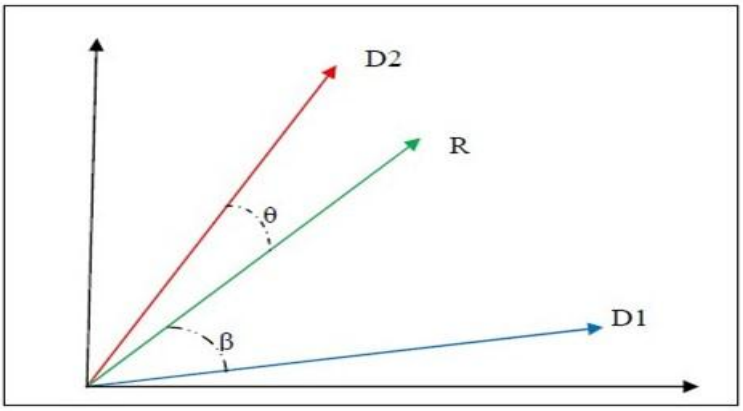


Figure 1.27 : Le modèle vectoriel.

Pour calculer la distance entre le vecteur du document D2 et le vecteur de la requête R, on doit calculer le cosinus de " θ " [BAZ 05] :

$$\text{Cos}\,\theta = \frac{D2 \cdot R}{\| D2 \| \cdot \| R \|}$$

Ou : R est le produit scalaire entre les deux vecteurs (D2 et R)

et ||D2|| est la norme du vecteur D2.

c. Le modèle probabiliste :

Le modèle probabiliste de recherche est basé sur le principe de la probabilité. Il consiste à classer les résultats selon la probabilité de pertinence d'un document par rapport à une requête. Les travaux les plus connus et les plus exploités dans ce modèle son ceux qui se basent sur la notion de "principe de classement probabiliste" qui indique que les documents sont classés selon leur probabilité de pertinence. Cette probabilité est estimée sur la base de la distribution des termes pertinents et non pertinent dans les documents:

Propose deux classes de documents D_j: pertinents (classe R) ou non pertinents (classe $\bar{R}$) par rapport à une requête Q :

- P(R/D) est la probabilité que D est pertinent pour Q
- P($\bar{R}$ /D) est probabilité que D n'est pas pertinent pour Q.

Pour calculer le score d'un document le modèle utilise la formule suivante :

$$S = \frac{P(R / D)}{P(\bar{R} / D)}$$

1.3.4. La recherche sémantique d'information :

Les recherches massives engagées ces dernières années, ont abouti à la réalisation de plusieurs systèmes qui considèrent la recherche sémantique d'informations sous divers angles, qui dépendent des buts envisagés, des stratégies et modèles utilisées. La plus part des modèles proposés partagent des aspects communs qui se concentrent sur les suivants points:

- ✓ des outils d'annotation.
- ✓ des représentations sémantiques des connaissances de domaines.
- ✓ des stratégies de recherche (sémantiques, hybrides, …. etc.) basées ou non sur les agents.
- ✓ des algorithmes de classements des résultats obtenus selon leur pertinence.

Le problème majeur dans les systèmes de recherche classiques est que l'utilisation des mots-clés simples comme requêtes conduit au problème de silence, où des documents pertinents ne seront pas présentés comme résultats de la requête. Les documents peuvent décrire le même contenu et le même sens en utilisant des termes déférents.

Les chercheurs ont pensé à utiliser la notion du concept, au lieu d'un simple terme. Un concept regroupe autour de lui plusieurs mots clés. Pour cela, les ressources sémantiques sont devenues un composant principal dans les moteurs de recherche sémantique (voir Figure 1.28).

La ressource sémantique, intervient pratiquement, dans toutes les étapes de la recherche d'information: dans l'indexation des documents, dans l'analyse de la requête et dans l'étape d'appariement requête-documents.

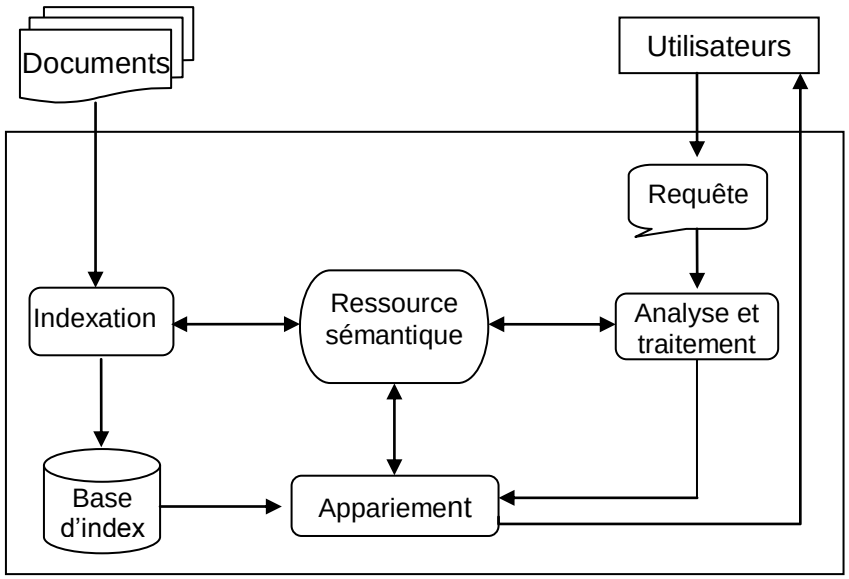


Figure 1.28 : Système de recherche sémantique de l'information [BOU 11].

a. Indexation sémantique :

Dans la recherche classique, l'unité textuelle dans l'indexation est représenté par un mot- clés. Cela pose deux problèmes : l'ambigüité et la disparité des mots.

- L'ambigüité sémantique due au fait qu'un terme peut avoir plusieurs sens selon le contexte ou il se trouve.
- La disparité des mots est dû au fait que les termes qui sont syntaxiquement différents ils ont le même sens.

Ces deux problèmes mènent à l'effet du silence où des documents pertinents ne sont pas retourné comme résultats pour une requête.

L'objectif de l'indexation sémantique est de désambigüiser le sens des termes. La tache consiste à représenter les documents et les requêtes par leurs sens. Dans ce type d'indexation l'unité textuelle sera le concept et non pas un simple terme.

Pour extraire le sens exact du terme, on doit utiliser les techniques de similarité sémantique pour trouver le meilleur concept qui représente le terme dans le contexte.

b. Traitement sémantique de la requête :

Le traitement sémantique de la requête d'un utilisateur peut résoudre deux problèmes, le premier est les erreurs dans l'orthographe de la requête par l'utilisation d'un dictionnaire et le deuxième est l'expansion automatique de la requête.

Pour l'expansion de la requête. il existe plusieurs méthode regroupé en deux techniques locale et globale. La méthode locale consiste à utiliser la notion de feedback, qui permet d'injecter de nouveaux termes dans la requête à partir des premiers documents retournés comme résultats.

La méthode globale permet de reformuler la requête en ajoutant d'autre termes proche sémantiquement aux termes initiaux de la requête (comme par exemple les synonymes...). ces termes sont calculés et extraits à partir d'une ressource sémantique.

c. Appariement sémantique :

L'appariement sémantique consiste à calculer la similarité sémantique entre les concepts de la requête et ceux des documents. Il permet de détecter les relations sémantiques entre les concepts, comme la relation de généralisation ou de spécialisation entre concepts. ces techniques permettent l'amélioration des résultats de recherche en retournant tous les documents qui ont une relation sémantique avec les concepts de la requête [BOU 11].

1.4. Conclusion :

Nous avons présenté dans ce chapitre un passage obligatoire sur les principes généraux du domaine qui englobe notre thème de recherche. On a débuté à définir le web sémantique et ces langages , puis on a enchainé par les ontologies avec un grand aperçu sur DBpedia , et pour finir, nous avons décrit la recherche d'information ainsi que les principales étapes d'un processus de recherche d'information .