# Deuxième partie

# Approche d'indexation sémantique à base de services web

# 3

# Conception des services web d'indexation

## 3.1 Introduction

Dans les chapitres précédents nous avons présenté les notions de base pouvant nous servir à mettre en œuvre une indexation sémantique à base de services Web. Dans ce chapitre, il s'agit en partie de mobiliser ces notions afin de concevoir des services web permettant une indexation sémantique de sources d'informations hétérogènes et distribuées de nature médicale. Ces services doivent, en outre, être capables de s'insérer dans le cadre d'un système de médiation (notamment lors de la réécriture de requêtes impliquant ces services d'indexation). De ce fait, la suite de ce chapitre sera consacrée à la présentation de notre proposition dans un cadre d'usage global dont l'objectif final est la médiation entre systèmes hétérogènes. Pour ce faire, nous allons commencer par décrire l'architecture globale du système de médiation. Cette architecture sera le

support du processus d'indexation et d'exploitation des résultats de notre approche. Nous allons décrire par la suite les principales étapes concernées par l'indexation à l'aide de diagrammes. Nous détaillerons également les caractéristiques des services web et les algorithmes nécessaires pour la réalisation de l'indexation.

## 3.2 Contexte général : Médiation à base de services Web

Dans la Figure 3.1, nous présentons l'architecture globale du système, afin de montrer la partie concernée par ce travail dans son contexte général, c'est à dire : l'architecture d'un système de médiation et d'interrogation des sources de données hétérogènes médicales (rapports médicaux, imagerie médicale annotée, ...) à base de services web.

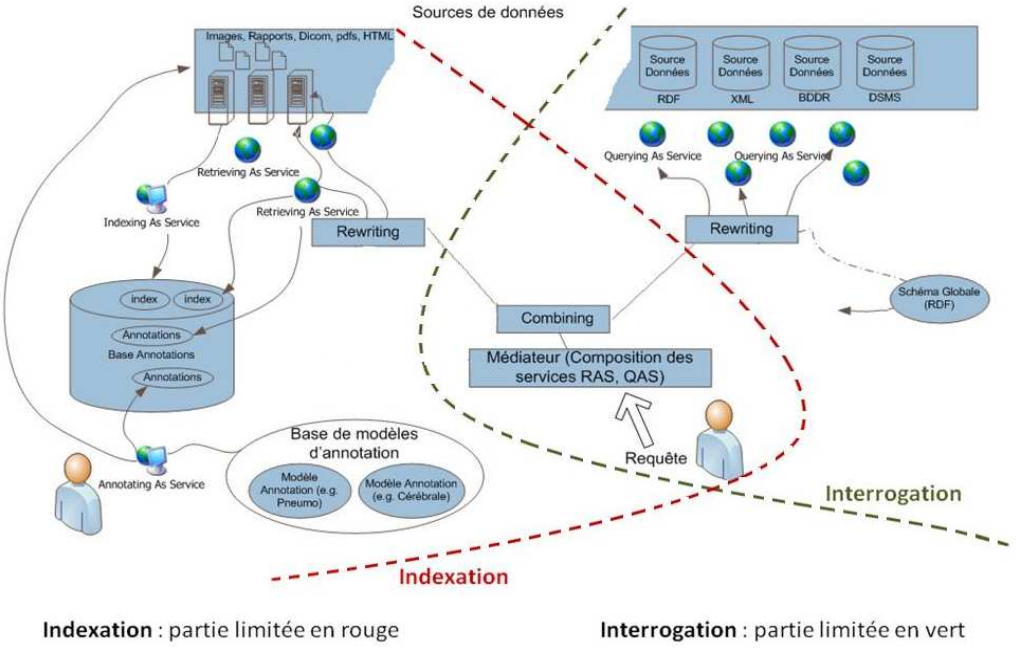


Fig. 3.1: Architecture du système global

### 3.2.1 Description de l'architecture globale

Le système global est divisé en deux parties, une partie pour l'interrogation et l'autre pour l'indexation. La partie interrogation concerne les sources de données structurées qui peuvent être interrogées par des langages classiques : SQL, SPARQL, ..., alors que la partie indexation concerne les sources de données semi ou non structurées qui peuvent être indexées et seront interrogées en utilisant des méthodes de recherche qui se basent sur des index.

La partie qui nous concerne dans ce travail est celle de l'indexation. Nous remarquons que dans notre partie indexation (délimitée en rouge) dans la Figure 3.1, l'utilisateur a comme vue l'interface du service web, qui se base sur un système intermédiaire représentant le système médiateur, c'est dans cette interface que l'utilisateur peut envoyer sa requête. Nous remarquons aussi qu'il peut y avoir une combinaison de services web.

### 3.2.2 Approche proposée

Bien que notre cadre global soit concerné par deux phases (l'indexation et l'interrogation moyennant des services web), notre étude se focalisera essentiellement sur le premier aspect. En effet, la phase d'indexation, étant en amont de l'interrogation, est au cœ œur de notre approche et en constitue l'élément premier pour la combinaison de différents services web pour l'évaluation d'une requête.

Notre objectif est de proposer des techniques d'indexation sémantique et de spécifier sous forme de services Web une interface permettant l'exploitation des index sémantiques proposés.

Le schéma de la Figure 3.2 représente l'approche proposée.

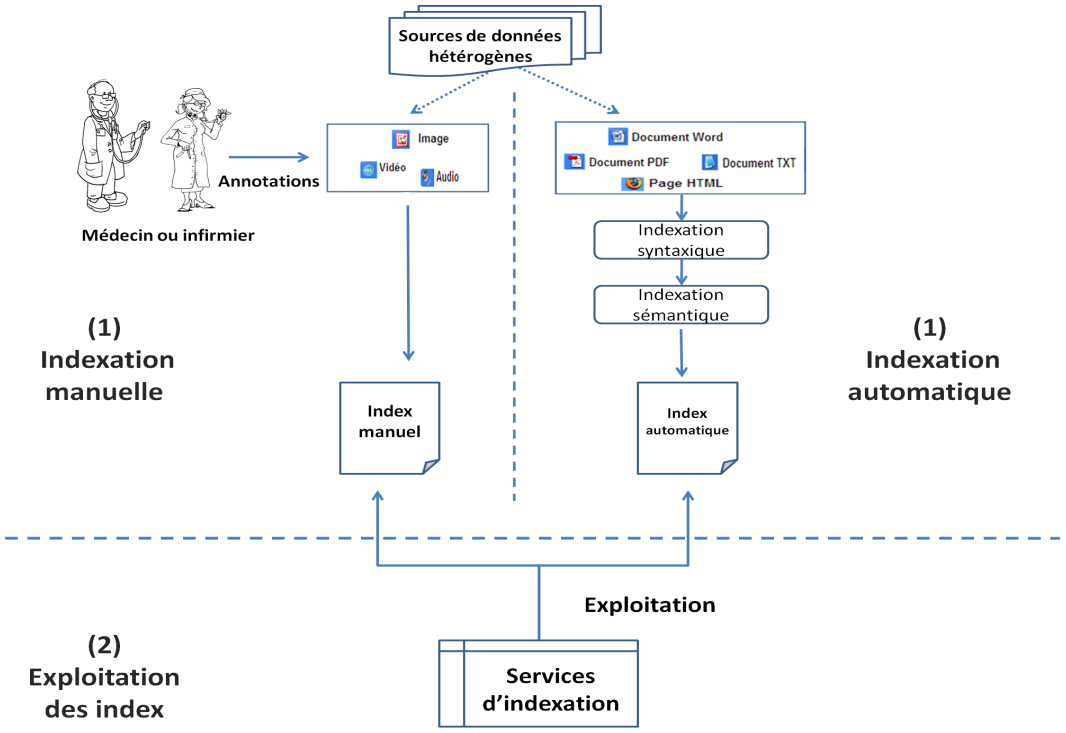


Fig. 3.2: Approche proposée

Sur le schéma de la figure précédente 3.2, nous remarquons que la phase d'indexation se réalise en deux traitements, selon le type de la source de données (sources de données hétérogènes) :

- Premier traitement : indexation faite automatiquement et concerne les sources textuelles (documents).
- Deuxième traitement : indexation faite d'une manière manuelle (des annotations faites par des humains) et concerne les sources non textuelles.

Les services web d'indexation considérés de ce travail concernent différents index issus de l'indexation faite préalablement. Nous distinguons notamment deux type index :

- index manuel = index issu des annotations de documents non textuels.
- index automatique = index issu d'une indexation automatique syntaxique et sémantique des documents textuels.

Ces services web d'indexation peuvent être intégrés et combinés avec d'autres services web d'interrogation (concernant les sources de données structurées) de l'architecture

globale, afin de trouver les sources de données pertinentes qui répondent à une requête donnée. Les réponses partielles obtenues par ces services sont combinées pour délivrer une réponse globale. Il est à noter que ces aspects concernant la réécriture de requêtes afin de d'obtenir une combinaison de résultats satisfaisant au plus une requête ne sont pas traités dans ce travail de recherche. Il sont traité dans le cadre d'un sujet de master recherche commencé parallèlement a notre travail.

## 3.3 Ontologie de médiation

Comme nous l'avons déjà mentionné, notre indexation s'inscrit dans le cadre d'une approche d'intégration de données par médiation [21]. Dans une telle approche, il est courant de définir, conceptuellement et de manière centralisée, un schéma global ou une ontologie regroupant l'ensemble des prédicats modélisant le domaine d'application du système médiateur. Dans notre cas qui est le domaine médical et afin de soutenir l'intégration des données des différentes sources, l'utilisateur posera ses requêtes dans les termes du vocabulaire structuré du domaine médical fourni par l'ontologie représentant l'ensemble des termes modélisés et utilisés par les différentes sources intégrées.

Le rôle de cette ontologie est d'établir la connexion entre les différentes sources accessibles en se fondant sur la définition de vues abstraites décrivant de façon homogène et uniforme le contenu des sources d'informations en termes des concepts de l'ontologie. Les sources d'informations pertinentes, pour l'évaluation d'une requête, sont calculées par réécriture de la requête en termes de ces vues (partie interrogation). Parmi ces vues, les services web d'indexation que nous allons proposer peuvent être utilisés.

### Un exemple d'ontologie médicale de médiation

Afin de montrer un scénario global dans lequel nous allons exemplifier notre approche d'indexation à base de services web, nous présenteront dans un premier temps le schéma global décrivant l'ontologie de notre système (Figure 3.3).

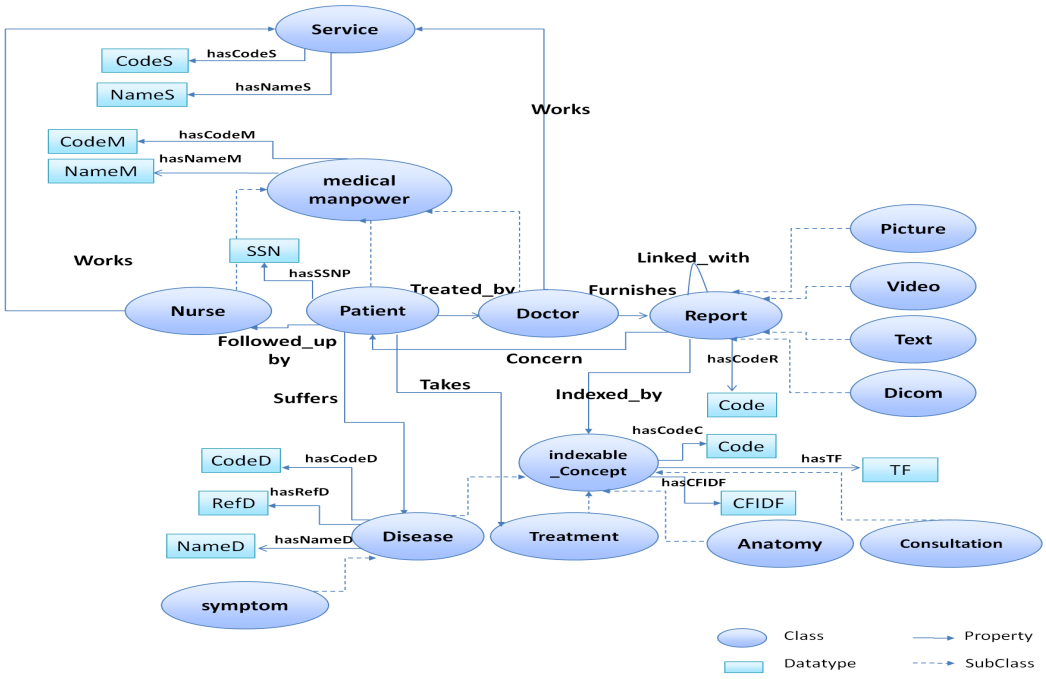


Fig. 3.3: Ontologie de médiation proposée

L'usage d'une ontologie lors de la phase d'indexation permet de rendre un certain nombre de services dont le plus important est la levée des ambiguïtés des sens des termes utilisés pour l'indexation. L'usage d'une ontologie permet aussi une meilleure représentation des connaissances contenues dans les documents. En termes d'indexation sémantique, les concepts de l'ontologie sont associés à chaque document selon les sémantiques qui y sont véhiculées. Ainsi, en plus de lier les documents à des termes pondérés comme dans les approches classiques [22], ces documents sont liés à des termes inter-connectés faisant partie d'une ontologie où les relations disposent d'une sémantique claire et non ambiguë (synonymie, équivalence, relation hiérarchiques, etc.). Dans cet exemple (Figure 3.3), notre ontologie du domaine d'application a été définie comme un ensemble de classes, et chaque classe dispose de :

- *propriétés*, e.g. la classe `Service` a un code de service (propriété : `hasCodeS`).
- *sous-classes*, e.g. le lien (`rdfs:subClass`) entre la classe `Doctor` et la classe `Medical_Manpower` signifiant que la classe `Doctor` est sous-classe de la classe `Medical_Manpower`.

Une classe peut être également liée à une ou plusieurs classes, e.g. `Report` est fourni

par un médecin `Doctor` et concerne un `patient`.

Afin de présenter les services Web pour l'indexation, nous allons présenter dans un premier temps les services web permettant l'interrogation des données dans le système de médiation. Ceci est nécessaire pour garder une cohérence notamment lorsque la réponse à une requête doit être fondée sur des services d'interrogation. Ces services d'interrogation doivent prendre en compte les caractéristiques de l'indexation pour permettre une future réécriture des requêtes et une combinaison des résultats.
Les différents services d'interrogation et d'indexation, seront décris par des vues RDF à partir de l'ontologie de médiation.

## 3.4 Services web d'interrogation

L'interrogation effective des sources se fait via un médiateur, qui traduit ou réécrit les requêtes en termes de vues. Comme nous l'avons mentionné précédemment dans ce mémoire, la partie interrogation des sources hétérogènes qui se base sur la réécriture des requêtes et de l'intégration des différents services web n'est pas notre objectif immédiat (c'est la seconde phase), néanmoins, il est nécessaire pour notre propos d'expliciter quelques exemples de services web d'interrogation afin de montrer comment ces derniers vont être combinés à nos services web d'indexation.
Le tableau suivant (Tableau 3.1) est un récapitulatif des différents services web d'interrogation utilisant la même ontologie proposée précédemment (l'ontologie de l'architecture globale).

| Service | Fonctionnalités | Contraintes |
|---|---|---|
| $S_1$($a, ?b) | Donne les médecins(b) traitant le patient(a) | $a < p1000$ |
| S'$_1$($a, ?b) | Donne les médecins(b) traitant le patient(a) | $a >= p1000$ |
| $S_2$($a, ?b) | Donne les infirmiers (b) qui soignent le patient(a) | |
| $S_3$($a, ?b) | Donne les médecins(b) travaillant dans un service(a) | |
| $S_4$($a, ?b, ?c, ?d, ?e) | Donne les rapports : image(b), vidéo(c), texte(d) et dicom(e) d'un patient(a) | |
| $S_5$($a, ?b) | Donne les maladies (b) traitées d"un patient (a) | |
| $S_6$($a, ?b) | Donne les infirmiers (c) travaillant dans un service(a) | |
| $S_7$($a, ?b) | Donne les patients (b) atteints d'une maladie (a) | |
| $S_8$($a, ?b) | Donne les rapports (b) fournis par un médecin (a) | |

Tab. 3.1: Services web d'interrogation

## Syntaxe utilisée

Le langage utilisé pour décrire les services web est RDF avec la syntaxe de représentation N3[1].

## Comment lire le service

Le symbole '$' représente les entrées (Inputs) et le symbole ' ?' représente les sorties (Outputs).

## Exemple d'utilisation

Le service web d'interrogation est une requête SPARQL conjonctive "contient l'opérateur and représenté par .". Prenons le service $S_1$ comme exemple. La définition de ce service en RDF avec la syntaxe N3 est la suivante :

```
S1($a,?b) :-
  (?M1  rdf:type   O:Doctor) .
```

[1]http ://www.w3.org/DesignIssues/Notation3

```
(?M1  O:CodeP  ?b) .
(?P1  rdf:type    O:Patient) .
(?P1  O:SSN_P   ?a) .
(?P1  O:Treated_by ?M1)
```

## Explications

$a (inputs) de type `Patient` et ?$b$ (output) de type `Doctor` dans notre ontologie.

- 'M1' est une variable de type 'Doctor' qui aura un 'CodeP' comme code personnel (du personnel médical) correspondant à chaque valeur de sortie qui est 'b'.
- 'P1' est une variable de type 'Patient' qui a comme code 'SSN_P' ayant comme valeur 'a'.
- Le patient 'P1' est traité par 'M1' (Treated_by).

On aura alors comme résultat : les médecins (b) traitant le patient(a).

Afin de schématiser graphiquement le service 'S1' représenté en RDF, on a la figure suivante (Figure 3.4) :

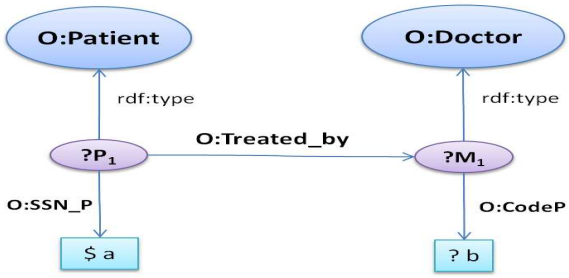


Fig. 3.4: Représentation RDF graphique du service S1 d'interrogation

### Description du sch$ma

Dans la représentation graphique de RDF, les nœuds représentent les variables, et les arcs sont des propriétés.

Ce service permet de trouver les Médecins correspondant à la variable $M1$ ayant un $CodeP$ correspondant à la variable $b$ en cherchant les patients correspondant à la variable $P1$ ayant un $SSN\_P$ correspondant à la variable $a$ tels que les patients sont traités par ces médecins (propriété $Treated\_by$).

## 3.5 Services web d'indexation

Pour notre système, nous avons proposé deux services web d'indexation décrits dans le tableau suivant (Tableau 3.2). Les services web d'indexation proposés exploitent les résultats de l'indexation automatique et manuelle.

| Service | Fonctionnalités | Contraintes |
|---|---|---|
| S1i($a,$seuil, ?b) | Donne tous les concepts indexables (b) cités dans le rapport (a) ordonnés selon n | $n >$ $seuil |
| S2i($a,$m, ?b) | Donne les $m$ premiers rapports (b) concernant un concept (a) ordonné par t, Rapport =texte et/ou vidéo et/ou dicom et/ou image | |

Tab. 3.2: Services web d'indexation

### Syntaxe utilisée

La syntaxe des services est celle de RDF.

### Explication des contraintes

1. Le service $S1i$ donne tous les concepts indexables (qui ont été indexés) du rapport (a). Ces concepts (les sorties) seront ordonnés d'une manière décroissante selon un certain $n$ ($n$ est le $CFIDF$ : propriété du concept qui sera définie plus tard) qui est une valeur entière calculée pour chaque concept, supérieur à un seuil donné en entrée également.
2. Le service $S2i$ donne les m premiers rapports (b) concernant un concept (a). Les rapports seront ordonnés aussi selon un $t$ ($t$ est le $tf$ (term frequency) : la fréquence d'apparition d'un concept dans un rapport, c'est une propriété du rapport qui sera détaillée par la suite). Les $m$ premiers de ces rapports seront pris comme résultat

## Exemple d'utilisation

Le service web d'indexation = requête SPARQL conjonctive augmentée de modificateurs de résultats (LIMIT, OFFSET, ORDER BY)

La définition du service S1i en RDF est la suivante :

```
S1i($a,$seuil, ?b) :-
    (?C1   rdf:type O:indexable\_Concept)  .
    (?C1   O:CodeC ?b) .
    (?C1   O:hasCFIDF ?CFIDF) .
    (?R1   rdf:type   O:Report) .
    (?R1   O:CodeR  ?a) .
    (?R1   O:Indexed_by  ?C1)
    FILTER (?CFIDF > $seuil)
    ORDER BY  ?CFIDF
```

## Explication

a (inputs) de type Rapport dans l'ontologie (Report) et b(outputs) de type Concept indexable, dans l'ontologie (indexable_Concept).

- '$C1$' est une variable de type `indexable_Concept` qui aura un `Codec` comme code correspondant à chaque valeur de sortie qui est $?b$. Chaque `C1` a une valeur `CFIDF`.
- Les `CFIDF` sont filtrés et seulement les concepts ayant le `CFIDF > seuil` seront pris dans les résultats.
- `R1` est une variable de type `Rapport` qui a comme code `CodeR` ayant comme valeur $a$ (l'entrée).
- Le rapport `R1` est indexé par `C1`.
- Les résultats seront ordonnés par `CFIDF`.

On aura alors comme résultat : les concepts (b) cités dans le rapport (a).

Afin de représenter graphiquement le service 'S1i' représenté en RDF, on a la figure suivante (Figure 3.5) :

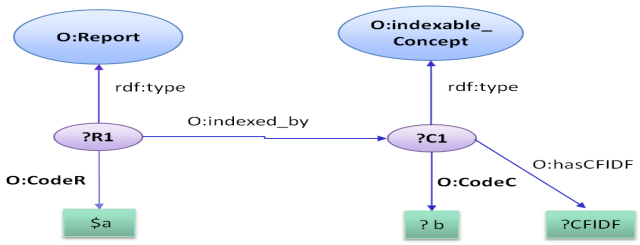


Fig. 3.5: Représentation RDF graphique du service S1i d'indexation

### Description du schéma

Ce service permet de trouver les concepts correpondant à la variable $C1$ ayant un $CodeC$. Les concepts trouvés doivent correpondre à la variable $b$ du rapport correpondant à la variable $R1$ ayant un $CodeR$, qui a $a$ comme variable. Les concepts indexent le rapport et sont ordonnés par la variable $CFIDF$.

Pour le second service, la définition RDF du service S2i est la suivante :

```
S2i($a,$m,?b):-
(?R1   rdf:type    O:Report) .
(?R1   O:CodeR  ?b) .
(?C1   rdf:type    O:indexable_Concept) .
(?C1   O:codeC   ?a) .
(?R1   O:indexed_by   ?C1) .
(O:indexable_Concept  O:hasTF ?TF) .
ORDER BY ?TF
LIMIT m
```

**Explication** : a (inputs) de type Concept indexable dans l'ontologie (indexable_Concept) et b (outputs) de type Rapport dans l'ontologie (Report).

- '$R1$' est une variable de type 'Rapport' qui aura un '$CodeR$' comme code correspondant à chaque valeur de sortie qui est '$b$'.
- '$C1$' est une variable de type 'Concept indexable' qui a comme code '$codeC$' ayant comme valeur '$a$'.

– Le rapport '$R1$' est indexé par '$C1$'.

– Les rapports (les sorties) seront ordonnés par '$tf$'.

– Les m premiers rapports seront pris uniquement.

On aura alors comme résultat : les rapports (b) concernant un concept(a).

La représentation graphique RDF correspondant à ce service 'S2i' est la suivante (Figure 3.6) :

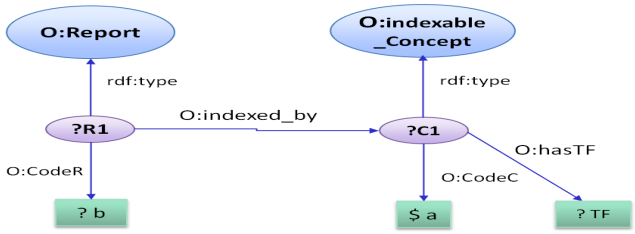


Fig. 3.6: Représentation RDF graphique du service S2i d'indexation

### Description du schéma

Ce service permet de trouver les rapports correpondant à la variable $R1$ ayant un $CodeR$ correpondant à la variable $b$ en cherchant les concepts correpondant à la variable $C1$ ayant un $CodeC$ correpondant à la variable $a$ tels que les rapports sont liés à/indexés par ces concepts par une propriété indexé_par (ObjectProperty) et sont ordonnés par la variable $TF$.

## 3.6 Exemple d'intégration des services web

Les services web d'indexation que nous avons proposé ne sont pas capables de résoudre tous les problèmes, la composition de services web permet de répondre aux besoins complexes des utilisateurs, par la combinaison de plusieurs services web. Dans l'exemple suivant, nous illustrons l'utilisation et l'intégration des différents services d'interrogation et d'indexation.

* Soit la requête de l'utilisateur suivante :

```
Donner les rapports contenant le concept 'y1' fournis par un médecin
travaillant dans un service 'ser0'./ y1 est une maladie par exemple.
```

### Requête en RDF

Dans ce qui suit nous présentons la requête en RDF :

```
Q(x1,y1,x2,y2):-
?R rdf:type  O:report
?R O:Indexed_by ?D
?R  O:CodeR   ?x1
?D rdf:type O:Disease
?D  O:hasNameD  ?y1
?R O:furnishes ?M
?M  rdf:type  O:Doctor
?M O:hasCodeM ?x2
?M O:hasNameM ?y2
?M O:dWorks ?S
?S  rdf:type O:Service
?S O:hasCodeS 'ser0'
?S O:hasNameS ?y3
```

### Représentation RDF graphique de la requête

La représentaion graphique de la requête en RDF est la suivante (voir Figure 3.7) :

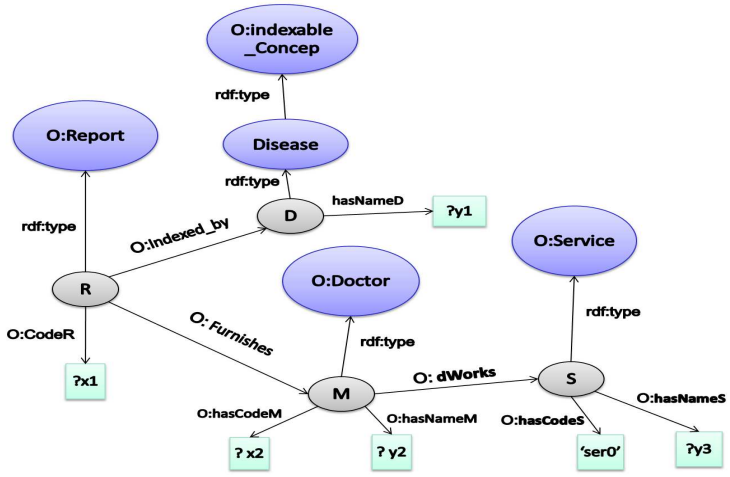


Fig. 3.7: Représentation RDF graphique de la requête

### Solution de la requête

La solution de cette requête est la suivante (Figure 3.8) :

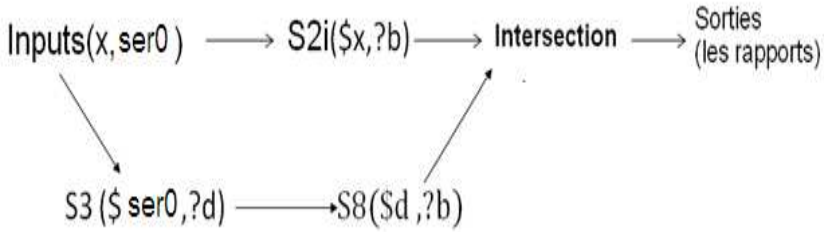


Fig. 3.8: Solution de la requête

Pour répondre à la requête, des services d'interrogation et d'indexation ont été intégrés.

Nous avons en entrée la maladie $x$ et le service $ser0$ :

1. Le service web d'interrogation S3 est lancé avec $ser0$ en entrée afin de trouver les médecins qui travaillent dans le service $ser0$.
2. Le service d'interrogation S8 est lancé après, avec comme entrée les médecins obtenus de (1) pour avoir les rapports fournis par ces médecins.
3. En parallèle à (1) et (2), le service web d'indexation S2i est lancé avec $x$ comme entrée qui est une maladie (c'est un concept car c'est une sous classe de 'indexable_Concept') afin de trouver les rapports contenant le concept $x$.
4. Une intersection est faite entre le résultat de (2) et le résultat de (3) pour avoir les rapports parlant de $x$ et fournis par les médecins travaillant dans $ser0$.

Dans le chapitre qui suit, nous présentons les étapes de notre approche d'indexation et de construction des index.