République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études

pour l'obtention du diplôme de Master en Informatique

Option: *Système d'Information et de Connaissances (S.I.C)*

Thème

La recommandation dans les Réseaux Sociaux avec l'utilisation des Données Liées

Réalisé par :

- **Mr. RAHILA Abdelkader**

Présenté le 21 Juin 2015 devant le jury composé de :

- *Mr. TADLAOUI Mohamed* *(Président)*
- *Mr. AMAR BENSABER Djamel* *(Encadreur)*
- *Mr. BENTAALLAH Mohamed Amine* *(Co-Encadreur)*
- *Mr. MATALLAH Hocine* *(Examinateur)*
- *Mme. HALFAOUI Amel* *(Examinateur)*

Année universitaire : 2014-2015

Remerciements

Je tiens d'abord à remercier profondément mes encadreurs Amar Bensaber Djamel et Mohamed amine Bentaallah de m'avoir accompagné durant cette aventure en guidant mes pas, en m'accordant leur confiance et en orientant notre collaboration durant ce projet. Sans leur implication, ce travail n'aurait jamais vu le jour.

Ma gratitude s'adresse à Messieurs TADLAOUI Mohamed, *MATALLAH Hocine et* Madame *HALFAOUI Amel* pour avoir acceptés de juger ce travail.

Ma gratitude s'adresse également à Mon père Allel de sa précieuse et sincère aide et conseils.

Je remercie aussi Monsieur Bouchikhi Smaine, de m'avoir aidé lors de mon séjour à Tlemcen.

Mes très chaleureux remerciements à mes collègues du master SIC, plus particulièrement messieurs MOULKHALOUA Ali et MAMMAR Osema.

Les mots ne suffisent pas pour remercier toute ma famille pour leur soutien. Ce travail est aussi le fruit de leur patience. Merci.

Table des matières

Liste des Figures

Introduction générale

1. Contexte

Nous vivons à l'ère de l'information. Au cours des dernières années, les services de réseaux sociaux ont gagné en popularité. Ils nous permettent une forte exploration et un partage de nos résultats de manière pratique.

Ces réseaux sociaux ont vu le jour au début des années 2000 avec l'arrivée du Web 2.0. Cette étape a permis aux utilisateurs de plus en plus nombreux d'accéder plus facilement à la publication de contenu via l'apparition des blogs et autres systèmes de collaboration tels que les forums de discussion ou les Wikis.

Cette transformation a été supportée par une transition technologique qui a vu les pages Web passer d'un format statique (HTML), à un format interactif grâce aux langages de programmation Web dit "dynamiques" tel que PHP, JSP ou ASP.

Les plates-formes des réseaux sociaux tels que Twitter, Facebook ont ouvert une nouvelle dimension à l'Internet en facilitant les interactions sociales.

Le Web sémantique fournit un modèle qui permet aux données d'être partagées et réutilisées entre plusieurs applications, entreprises et groupes d'utilisateurs. Le Web sémantique propose des langages spécialement conçus pour les Données : RDF (Resource Description Framework), OWL (Web Ontology Language), et XML (extensible Markup Language). L'utilisation de ces langage a permis de décrire des choses telles que des personnes, des réunions, photo, … etc. étant donné que le HTML ne décrivait que les documents et les liens entre eux.

L'apparition **des données liées** (Linked Data) a donné un essor pour le web. Cette technologie est très importante et utile pour découvrir des nouvelles informations, d'y accéder, de les intégrer et de les utiliser dans des applications grâce à l'utilisation des liens RDF externes qui permet de rendre le web un espace de données interconnectés.

2. Problématique

Nous sommes partis du constat que les utilisateurs des réseaux sociaux se retrouvent confrontés en permanence à un véritable déluge d'informations amplifié par le phénomène du Web social ces dernières années. Nous nous intéressons au défi de fournir à l'utilisateur une expérience de qualité sur le Web social. Afin de parvenir à nos fins, nous explorons deux approches complémentaires.

Dans un premier temps, il s'agit de fournir un mécanisme d'enrichissement du profil de l'utilisateur Twitter en utilisant des données publiées en qualité Linked Data afin de résoudre le problème de la détermination des nouveaux centres d'intérêts.

En second lieu, il s'agit de s'attaquer au problème de la découverte d'informations sur le réseau en fournissant des recommandations de livres. La tâche consiste en l'occurrence à fournir à l'utilisateur une liste de livres correspondante à ses centres d'intérêts.

3. Contributions

L'objectif de ce projet est d'étudier les challenges liés à la forte utilisation du web social et de l'apport du web sémantique et du Linked Data pour mieux appréhender les problématiques des réseaux sociaux.

Nous nous intéressons particulièrement au cas de Twitter qui est un réseau social numérique très populaire, disposant d'une part une API qui est plus riche en termes de fonctionnalités et la plus utilisée par les développeurs d'autre part.

Dans notre contribution, nous avons en premier temps pris la préoccupation de construire un profil utilisateur dans le réseau social numérique et d'essayer en deuxième temps de l'enrichir avec un jeu de données du web plus important publiées en qualité Linked Data afin de construire les nouveaux centres d'intérêts de l'utilisateur. Et en dernier lieu on a proposé des recommandations de livres en se basant sur ses centres d'intérêts.

4. Organisation du mémoire

Ce mémoire est organisé comme suit :

Après l'introduction qui présente le contexte du projet, la problématique étudiée et notre contribution, nous présentons le premier chapitre, qui parle des notions et principes du web social.

Le deuxième chapitre, s'intéresse au Web sémantique ainsi que ses différentes technologies et mets l'accent sur le web social sémantique.

Le troisième chapitre mets la lumière sur les données liées, ainsi que leurs utilités dans le cadre du web social.

Le quatrième chapitre est consacré aux travaux similaires existant dans le domaine de recommandation.

Le cinquième chapitre « conception et réalisation » représente l'architecture de notre approche, les technologies et les outils utilisés pour l'implémentation et la mise en œuvre de notre application qui repose sur les points suivants :

- Analyse du réseau social Twitter en traitant les activités des utilisateurs ainsi que les activités des amis de son réseau egocentrique.
- Enrichissement du profil obtenu en exploitant les données liées du web et plus spécialement DBLP.
- Recommandation de livres en utilisant des requêtes SPARQL pour interroger des données publiées en qualité Linked Data.

Et enfin, nous clôturons ce rapport par une conclusion générale tout en ouvrant une fenêtre sur les futurs travaux dans le contexte de ce projet de fin d'études.

Chapitre 1 Web Social

1.1. Introduction

Depuis leur introduction, les sites des réseaux sociaux tels que *Myspace*, *Facebook*, Twitter, *Cyworld*, et *Bebo* ont attiré des millions d'utilisateurs, dont beaucoup d'entre eux ayant intégré ces sites dans leurs pratiques quotidiennes. A ce jour, il y a des centaines de sites de réseautage personnel, avec différents apports technologiques, pour soutenir un large éventail d'intérêts et de pratiques. Alors que leurs caractéristiques technologiques clés sont assez uniformes, les cultures qui émergent autour des SRS sont variées, également dans la mesure où ils intègrent de nouvelles informations et des outils de communication, tels que la connectivité mobile, *les blogs*, et *photo*, *vidéo-partage*. [Danah m. boyd2007]

Le Réseaux sociaux étudient et analysent les structures des relations liant des individus (ou d'autres groupes sociaux, comme les organisations) et l'interdépendance des comportements ou des attitudes liées à des configurations de relations sociales [A. James O'Malley Æ Peter V. Marsden, 2008]

Les médias sociaux tels que *Facebook* et *Flickr* sont devenus une plate-forme très populaire de partage de contenus multimédia, avec plus de 40 millions de photos mensuels ajoutées dans *Flickr* et plus de 200 millions de photos envoyées par jour dans *Facebook* qui compte actuellement près de 90 milliards de photos au total. Si l'utilisateur veut recueillir tous les éléments multimédias d'un événement social spécifique, cette tâche devient très difficile [Tim O'Reilly 2005]

Les médias sociaux diffèrent dans le traitement des contenus stockés. Certaines d'entre elles préservent les métadonnées des contenus tels que les métadonnées géo-temporelle comme dans le cas de *Flickr*, d'autres, comme le cas de *Facebook* enlèvent la plupart de ces métadonnées incluses dans l'en-tête EXIF, mais ils conservent l'information qui est ajoutée manuellement par l'utilisateur pour les albums.

1.2. Les réseaux sociaux

Un réseau social est une structure comportant un ensemble d'acteurs qui sont impliqués sur certains types d'interactions. Un acteur est une entité sociale qui pourrait être une seule personne, un groupe ou une entreprise. Les acteurs sont reliés les uns aux autres par des liens qui peuvent désigner une ou plusieurs relations. Ces liens peuvent être de différents types, y compris des liens d'amitié, des liens de collaboration, des liens d'affaires, etc. Par conséquent, on peut distinguer :

- **Les réseaux hétérogènes :** des réseaux sociaux où plusieurs types d'acteurs ou plusieurs types de liens peuvent exister (par exemple, les acteurs des réseaux sociaux liés à différents types de liens, c'est à dire, les collègues et les amis).

- **Les réseaux homogènes :** ce sont les réseaux sociaux où il existe un seul type d'acteur avec un seul type de lien entre les acteurs (par exemple, les acteurs des réseaux sociaux connectés en utilisant uniquement des liens d'amitié).

Dans la pratique, les réseaux sociaux offrent aux internautes de nouveaux moyens et façons de se connecter, de communiquer et de partager des informations avec d'autres membres au sein de leurs plates-formes intéressantes. En théorie, ces réseaux sociaux sont composés de plusieurs éléments, peuvent contenir différents types de données, et avoir différentes représentations.

1.3. Le Web Social

« *Le Web est une création sociale plutôt que technologique .Je l'ai créé en envisageant un effet social pour faciliter la collaboration entre des personnes - ce n'est pas un jouet technique.* » affirme Tim Berners-Lee dans son livre "Weaving the Web" [Berners-Lee 2000]. Malgré cette intention dans les premières années du Web il était difficile de se servir du Web pour se communiquer. Le contenu du Web était créé principalement par une minorité d'utilisateurs formés en HTML, ce qui rendait le Web un canal de diffusion non participatif.

C'est avec l'émergence des blogs, wikis, forums et d'autres sites destinés à la collaboration que le Web a commencé à prendre la forme d'un outil d'intégration dans l'aspect social du Web. Ce phénomène et son impact sur l'industrie du Web et sur l'apparition de nouveaux services innovants ont notamment été repérés par O'Reilly en 2005[Tim O'Reilly 2005]. Deux éléments clés ont permis au Web de devenir plus interactif et plus participatif :

a. La facilité de création du contenu : De nombreux sites ont proposé des pages Web qui permettaient à l'utilisateur de fournir ses propres textes, documents, images, vidéos,… etc., de les incorporer au Web très facilement, sans devoir rédiger de code HTML. Ce sont des composants de ces nouveaux sites, « appelés applications Web», qui se chargent de transformer les données de l'utilisateur aux formats du Web. Le contenu généré par les utilisateurs était à l'origine d'une nouvelle excroissance du Web, qui persiste encore au moment de l'écriture de ce mémoire. Parmi les nombreux types de sites sociaux nous pouvons distinguer notamment :

- **Blogs :** sites qui permettent à chacun d'avoir son propre journal (publication) sur le Web. Très populaire à l'époque de la naissance du Web Social, cette Pratique évolue de plus en plus vers la publication de billets très courts, reconnue sous le nom «*Microblogging*» ; Nous appelons ces billets courts (d'une taille souvent limitée à 140 caractères) : les *tweets*. Le b*log* est un outil de publication, d'échange, et de partage du contenu dans le web.

- **Wikis :** site permettent à un groupe de personnes de développer un site Internet de manière *collaborative* alors qu'ils n'ont aucune notion de HTML ou autre langage de programmation. N'importe qui peut modifier les pages. Le wiki le plus connu est l'encyclopédie en ligne : *Wikipédia*.

- **Sites de partage de contenu :** tels que *Flickr* et *YouTube* permettent aux utilisateurs d'afficher leurs collections photographiques et leurs vidéos et de les exposer aux commentaires de la communauté en ligne.

- **Les agrégateurs d'actualité** comme *Digg* permettent aux internautes de partager des actualités qu'ils ont trouvées en ligne, de les commenter ou voter pour les contenus préférés. Les éléments les plus populaires passent en première page des sites pour les faire connaître au plus grand nombre.

- **Les sites de favoris sociaux** (ou d'étiquetage) comme *Delicious, Blogmarks* et *StumbleUpon* permettent aux utilisateurs d'étiqueter, d'enregistrer, de gérer et de partager des contenus web. Ces sites permettent d'enregistrer des sites web favoris, puis de les classer par thèmes et mots-clés.

- **Forums** : site avec des espaces d'échanges dédiés. Les discussions y sont archivées ce qui permet une communication asynchrone pour faciliter la discussion au sein des communautés en ligne.

- **Réseaux Sociaux** : sites destinés à la socialisation et la mise en relations des individus permettant les échanges de différents éléments (ex. messages, photos, liens, commentaires) comme *Facebook, Twitter, LinkedIn* ou *Viadeo*.

D'autres outils tel que *Ning*, offrent aux internautes la possibilité de créer leur propre réseau.

- **De nombreuses applications pour Smartphone** (*Four square*, *Instagram*, *Yelp, Path*…) proposent également des fonctions communautaires notamment pour se localiser dans un lieu, donner un avis (restaurant, musée, commerce…), pour prendre des photos géo localisées puis les partager avec le réseau d'utilisateurs de cette même application et/ou avec son propre réseau *Facebook* et/ou *Twitter*.

b. Rechargement des pages en temps réel : Les techniques pour l'affichage des parties de pages Web, telles qu'Ajax (Asynchronous JavaScript and XML) ont commencé à émerger vers 2005[Garrett, J 2005]. Ces techniques permettent de mettre à jour certains éléments de pages Web, sans même l'intervention de l'utilisateur, créant ainsi un effet de dynamique et de continuité. Un bon exemple de cette pratique est le site *Twitter* qui l'utilise pour donner l'impression d'un flux de messages qui passent en temps réel devant les yeux de l'utilisateur.

Si cette nouvelle facilité d'interaction avec le Web a rendu plus fréquents les échanges entre les utilisateurs, elle a également fait émerger un nouveau phénomène - la socialisation centrée sur l'objet - repéré d'abord par Engestrom [Engestrom, J 2005] et puis étudié par Breslin et Decker [Breslin, J., Decker, S.2007]. En interagissant avec des objets Web, les utilisateurs laissent des traces, par exemple sous forme de commentaires visibles par d'autres utilisateurs interagissant avec le même objet. Cette pratique peut conduire à des conversations et à l'établissement de nouveaux contacts. La socialisation autour du contenu Web est donc à la fois la cause et la conséquence de la création de contenus par des utilisateurs.

Il est important de souligner deux tendances qui émergent incontestablement sur le Web Social et qui transforment de manière significative la nature des éléments du Web :

a. La publication des éléments Web de petite taille, appelés *microposts*. Les sites Web proposent de plus en plus des interactions nécessitant peu d'effort de la part de l'utilisateur. Par exemple sur le réseau social *Facebook* l'utilisateur peut apprécier un objet Web en cliquant sur un bouton «*j'aime*» qui y est associé .Ces actions résultent de la création des nombreux micro-objets et intensifient l'interaction entre le Web et l'utilisateur. Un autre réseau social, *Foursquare* permet aux utilisateurs de se déclarer présents dans un lieu localisé par *GPS* en effectuant un «*Check-in*» qui sert à signaler leur présence dans ce lieu en temps réel à leurs contacts. L'émergence de tweets courts, faciles à créer, publier et republier fait également preuve de la réalité de cette tendance.

.

b. **L'émergence du contenu en tant que conséquence indirecte d'une action de l'utilisateur :** Certaines applications Web, de plus en plus nombreuses, identifient les activités de l'utilisateur et génèrent du contenu Web, sans l'intervention de l'utilisateur. C'est le cas de *Spotify*, le service de streaming musical, qui publie sur *Facebook*, de manière automatique l'information sur chaque chanson que l'utilisateur écoute.

Nous pouvons définir les médias sociaux selon les 5 piliers suivants : [Thierry Wellhoff 2012]

-Participation : Tout est fait pour encourager les internautes à contribuer et donner leur avis, supprimant ainsi la barrière entre publics et médias.

-Ouverture : Les médias sociaux se basent sur les principes de collaboration et d'échange d'informations. Tout le monde peut y prendre part, il n'y a aucune barrière à l'entrée.

-Conversation : Alors que les médias traditionnels ont tendance à «raconter» ou à transmettre un message, les médias sociaux sont plus dans le dialogue, ce qui implique une écoute attentive.

-Communauté : Les médias sociaux permettent de constituer rapidement des communautés de personnes partageant les mêmes intérêts.

-Interconnexion : La plupart des médias sociaux se développe par interconnexion en tirant partie des liens avec les autres sites, ressources ou personnes.

1.4. Développement historique du Web social

Le web des années 1990 ressemblait beaucoup à la combinaison d'un annuaire téléphonique et les pages jaunes et malgré la puissance de raccordement de liens hypertextes, il donne peu de sens à la communauté parmi ses utilisateurs. [Berners-Lee 2000]

Cette attitude passive à l'égard du Web a été brisée par une série de changements dans les habitudes d'utilisation et technologiques qui sont désormais désignés comme Web 2.0, un mot inventé par Tim O'Reilly [Tim O'Reilly]. Dans ce qui suit, nous résumons l'histoire et les caractéristiques qui définissent la Web 2.0. Les changements qui ont conduit à son niveau actuel d'engagement social en ligne n'ont pas été radicales ou individuellement significatifs. Néanmoins, ce jeu d'innovations dans l'architecture et les modèles d'utilisation du Web a conduit à un rôle tout à fait différent du monde en ligne comme plate-forme pour la communication et l'interaction sociale intense.

L'augmentation de notre capacité à obtenir de l'information et du soutien social en ligne peut être quantifiée. Une récente enquête d'envergure basée sur des entretiens avec 2200 adultes montre que l'Internet améliore de manière significative la capacité de maintenir leurs réseaux sociaux en dépit de craintes initiales concernant les effets de la diminution du contact avec la vie réelle. L'enquête confirme que non seulement les réseaux sont maintenus et étendus en ligne, mais ils sont également activé avec succès pour traiter les situations de la vie tels que l'obtention d'un soutien en cas de maladie grave, à la recherche d'emplois, et de s'informer sur les grands investissements, etc... La première vague de socialisation sur le Web était due à l'apparition des blogs, des wikis et d'autres formes de communication et de collaboration en ligne.

Les premiers réseaux sociaux en ligne (également appelés services de réseaux sociaux) entrés sur le terrain en même temps que les blogs et les wikis ont commencé à décoller. En 2003, le premier arrivant Friendster25 attiré plus de cinq millions d'utilisateurs enregistrés en l'espace de quelques mois, qui a été suivi par Google et Microsoft de départ ou d'annoncer des services similaires. Bien que ces sites disposent d'une grande partie de la même teneur qui apparaissent sur les pages Web personnelles, ils fournissent un point d'accès central et ajoute la structure dans le processus de partage des renseignements personnels et de la socialisation en ligne.

En termes de mise en œuvre, les nouveaux sites web se fondent sur de nouvelles façons d'appliquer certaines des technologies préexistantes. (Asynchronous JavaScript et XML ou AJAX, ce qui entraîne la plupart des derniers sites Web est simplement un ensemble de technologies qui ont été prises en charge par les navigateurs depuis des années). Ce qui peut être également observé une préférence pour les formats, les langues et les protocoles qui sont faciles à utiliser et à développer avec, en particulier des langages de script, de formats tels que JSON, des protocoles tels que REST. Il s'agit de soutenir le développement et le prototypage rapide (Flickr, par exemple, est connu pour adapter l'interface de l'utilisateur plusieurs fois par jour). [Amit Sheth Ramesh Jain2007]

1.5. Le réseau traditionnel et le réseau social en ligne

Les individus se regroupent sur la base d'intérêts communs ou de valeurs partagées et le modèle de réseautage traditionnel s'est tout simplement transposé sur le Web. L'abolition de limites géographiques, temporelles et, jusqu'à un certain point, psychologiques semble être un facteur déterminant dans la propulsion du réseautage en ligne. Voici un aperçu des différences identifiées entre le réseau traditionnel et le réseau en ligne.

Réseau traditionnel	Réseau social en ligne
Selon une base géographique.	Sans frontières.
Basé sur des intérêts communs.	Basé sur des intérêts communs.
Limité par la classe sociale.	Sans limites.
Diffusion restreinte de l'information.	Diffusion en temps réel de l'information.
Pouvoir des leaders d'opinion limité à une présence dans les médias traditionnels ou à des actions en personne.	Présence des leaders d'opinions en ligne très importante. Influence en temps réel et exponentielle.
Diffusion et promotion de l'innovation et des nouveautés limitée par les lieux physiques ou par les médias traditionnels nécessaires à la communication.	Diffusion et promotion de l'innovation et des nouveautés en temps réel.
Information personnelle inexistante ou limitée au groupe d'appartenance.	Affichage en ligne d'information personnelle sur les membres.

Tableau 1 Comparaison entre le réseau traditionnel et le réseau social en ligne

Ce tableau permet de constater deux grandes différences entre les réseaux traditionnels et les réseaux sur le Web. D'une part, le réseau en ligne est caractérisé par la notion d'instantanéité. D'autre part, il démontre la grande ouverture causée en partie par la réduction des limites physiques.

1.6. Analyse des réseaux sociaux

Pour analyser, étudier, extraire des informations à partir de réseaux sociaux, on a deux catégories principales :

a. SNA : a été développé par les sociologues à découvrir les propriétés des réseaux sociaux en mettant l'accent sur les relations sociales entre les acteurs d'un réseau. Plusieurs études ont été menées depuis les années 1930 [S. Wasserman & K. Faust 1999] l'exploitation de structures, d'identifier les acteurs de réseaux de liens mondiaux des rôles et des positions, et en examinant les modes d'interaction au sein des réseaux sociaux. Bien que les techniques d'analyse de réseaux sociaux révèlent des informations importantes sur les réseaux sociaux, leur principal objectif est de mesurer les propriétés structurelles des réseaux par l'étude des réseaux «topologies ».

b. **Link Mining** : a été introduit par des informaticiens pour extraire des motifs cachés à partir des données disponibles. Elle peut être considérer comme la tâche d'appliquer des techniques d'exploration de données sur les réseaux, tout en tenant compte explicitement sur les liens entre les acteurs des réseaux sociaux [L. Getoor and C.P.Diehl 2005]. L'objectif de l'exploration de données est de trouver des connaissances inconnues cachées et potentiellement utiles à partir d'une grande quantité de données. En fait, les techniques d'exploration de données sont devenues vitales pour découvrir des informations cachées de réseaux sociaux.

Le Web 2.0 a maintenant émergé comme le principal concurrent de la prochaine évolution du Web. Les chercheurs, les développeurs sont tous affluent vers la bannière du Web 2.0, basé sur sa promesse d'une augmentation massive de partage et de participation parmi les internautes [P. Mika, M. Greaves 2008], L'accent principal de services produits était sur le contenu social et généré par les utilisateurs. Ceci a été réalisé en mettant un ensemble des technologies différentes lisible par machine formats tels que XML (Atom, RSS, etc.) [Romain Blin,Julien Subercaze Henry Story 2012]

1.7. Conclusion

L'émergence des réseaux sociaux est liée aux révolutions technologiques et techniques. L'apparition de la technologie AJAX (JavaScript + XML) a permis des interactions plus rapides avec les pages Internet. De ce fait, le nombre de membres de ces réseaux sociaux s'est allongé. D'une part car les interactions étant plus rapides, consulter Internet est devenu plus confortable. Mais d'autre part, car les utilisateurs prennent conscience de leur pouvoir d'interagir sur la toile. C'est ce qui a donné naissance au Web 2.0.

Les inconvénients de ces multiples interactions sont la désorganisation des données. Intervient alors un concept, celui des métadonnées, qui vont permettre de garder des interactions tout en structurant les données (c'est le web sémantique). Ce concept marque la naissance du Web 3.0. Les données sont alors plus facilement exploitables par nos outils.

Chapitre 2
Web Social sémantique

2.1. Introduction et définition

L'idée d'un Web sémantique a été introduite par Tim Berners-Lee dans les années quatre-vingt-dix. Une dizaine d'années plus tard, une définition simple et claire du Web sémantique également connu sous le nom de Web 3.0 a vu le jour.

Amit Sheth Ramesh Jain [Amit Sheth Ramesh Jain 2007] a décrit que le Web sémantique apportera de la structure au contenu des pages Web. Du point de vue de leur sens, en créant un environnement où les agents logiciels voyageant de page en page pourront facilement effectuer des tâches complexes pour les utilisateurs. Le défi du Web sémantique est donc de fournir un langage qui exprime à la fois des données et des règles de raisonnement sur ces dernières, et qui permet d'exporter sur le Web des règles de tout système de représentation des connaissances.

Le Web sémantique permettra aux machines de comprendre des documents et des données sémantiques. Le sens est exprimé en RDF (Resource Description Framework) et encodé en ensembles de triplets, chaque triplet étant comme le sujet, le verbe et l'objet d'une phrase élémentaire. Ces triplets peuvent être écrits à l'aide de balises XML. En **RDF**, un document effectue des affirmations selon lesquelles des choses particulières (des personnes, des pages Web ou n'importe quoi d'autre) possèdent des propriétés (comme "*est une sœur de*", "*est l'auteur de*") avec certaines valeurs (une autre personne, une page Web). Cette structure se révèle être un moyen naturel de décrire la grande majorité des données traitées par des machines. Un sujet et un objet sont chacun identifié par un Universel Resource Identifier (URI), de la même manière que l'on utilise un lien sur une page Web. [Luciano Floridi 2009]

2.2. Le langage de représentation des connaissances RDF

Ce qui rend la connaissance représentée sur le web différente de la représentation des connaissances classiques selon Tim Berners-Lee, comme indiqué dans la première conférence du World Wide Web à Genève en 1994, était que l'ajout de la sémantique au Web implique deux choses [Romain Blin & ALL 2012]

Le premier point, c'est que les documents doivent disposer d'informations sous des formes lisibles par la machine .Telles informations nécessitent un langage de représentation des connaissances qui a une sorte de langage relativement neutre quant au contenu pour l'encodage.

Le deuxième point, et de permettre de créer des liens avec des valeurs relationnelles. Cela signifie que le modèle de base du Web sémantique sera un reflet du Web lui-même.

Le Web sémantique consiste à relier les ressources par des liens. Le Web sémantique est alors facilement interprété comme un descendant des réseaux sémantiques et de l'intelligence artificielle classique (IA), où les nœuds sont les ressources et les arcs sont des liens. Le premier langage de représentation des connaissances pour le Web sémantique, est le Resource Description Framework (RDF) recommandé par W3C.

2.2.1. RDF et le principe d'universalité

Selon le principe d'universalité, du fait qu'une ressource peut être n'importe quoi, alors le langage de représentation des connaissances devrait la considérer comme une ressource et un URI doit lui être attribué. Au lieu d'étiqueter les arcs et les nœuds avec des termes du langage naturel, en RDF tous les arcs et les nœuds peuvent être étiquetés avec les URI.

Cela a des avantages, car RDF permet d'être utilisé pour modéliser les relations entre les ressources accessibles sur le Web et même mélanger certains types de relations. Cette sorte de «métadonnée» est illustrée par la relation entre une page web et son auteur humain, dans laquelle l'auteur et la page seraient désignés par les URI en utilisant RDF qui peut alors modéliser ses propres constructions de langage en utilisant des URI, et faire des déclarations au sujet de ses propres constructions de langage de base. Cependant, toutes les composantes de RDF peuvent être considérées comme des ressources, de même que toutes les ressources ne peuvent pas avoir les URI. Par conséquent toutes les composantes du RDF peuvent ne pas avoir les URI. Par exemple, une chaîne de texte ou un numéro peut être un composant de RDF, et ceux-ci sont appelés littéraux RDF.

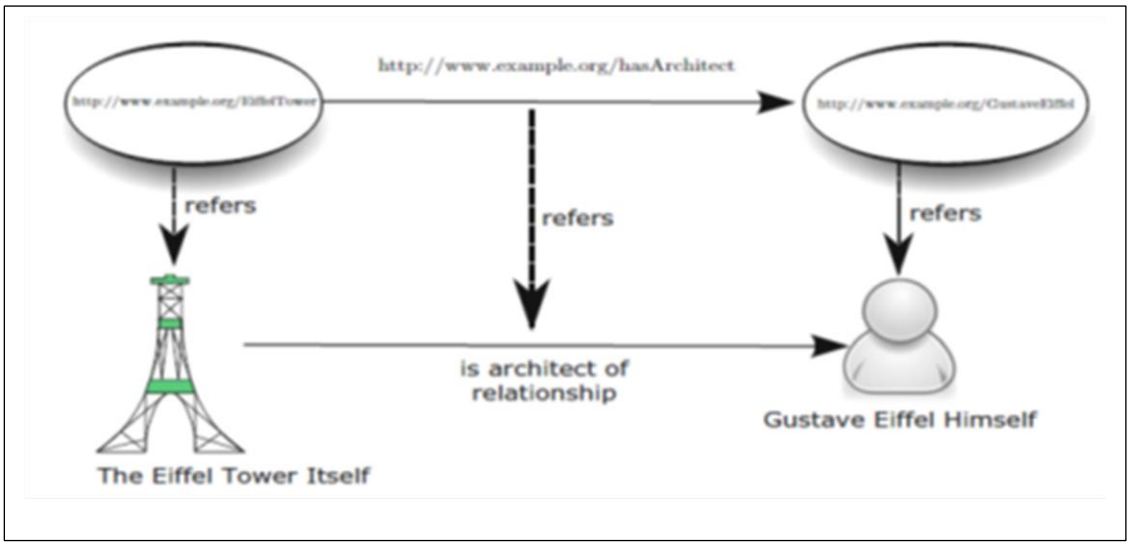

Figure 1 exemple RDF

2.2.2. RDF et le principe de lien

La deuxième étape dans la vision de Tim Berners-Lee pour le web sémantique « permettant de créer des liens avec des valeurs de la relation», qui suit simplement de l'application du principe de l'universalité de la représentation des connaissances.

Depuis RDF est composé de ressources, et de toute ressource on peut créer un lien vers une autre ressource, puis toute expression en RDF peut être lié à un autre terme. Ce lien constitue le cœur du RDF, car elle permet aux URIs d'être reliés entre eux pour une déclaration en RDF. La forme précise de la déclaration en RDF est un triplet qui est constitué de deux ressources reliées par un lien, comme le montre la *Figure 1*.

La Représentations Web dans une certaine forme de langage du Web sémantique comme RDF sont appelés documents du Web sémantique. Il y a plusieurs options pour le codage des documents du Web sémantique. L'encodage standardisé W3C RDF est le format RDF / XML, même si un codage plus simple appelée Turtle existe.

RDF est un format de données abstrait qui peut être écrit dans plusieurs sérialisations. Une des sérialisations est XML, le langage de balisage extensible, qui est représenté en (Figure 2)

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF [
<!ENTITY bbca "http://www.bbc.co.uk/music/artists/">
<!ENTITY bbci "http://www.bbc.co.uk/music/images/artists/">
<!ENTITY mba "http://musicbrainz.org/artist/">
]>

<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:owl="http://www.w3.org/2002/07/owl#"
 xmlns:foaf="http://xmlns.com/foaf/0.1/"
 xmlns:mo="http://purl.org/ontology/mo/">

 <mo:MusicArtist rdf:about="&bbca;d87e52c5-bb8d-4da8-b941-9f4928627dc8#artist">
   <rdf:type rdf:resource="http://purl.org/ontology/mo/MusicGroup"/>
   <foaf:name>ABBA</foaf:name>
   <foaf:homepage rdf:resource="http://www.abbasite.com/"/>
   <mo:image rdf:resource="&bbci;542x305/d87e52c5-bb8d-4da8-b941-9f4928627dc8.jpg"/>
   <mo:member rdf:resource="&bbca;042c35d3-0756-4804-b2c2-be57a683efa2#artist"/>
   <mo:member rdf:resource="&bbca;2f031686-3f01-4f33-a4fc-fb3944532efa#artist"/>
   <mo:member rdf:resource="&bbca;aebbb417-0d18-4fec-a2e2-ce9663d1fa7e#artist"/>
   <mo:member rdf:resource="&bbca;ffb77292-9712-4d03-94aa-bdb1d4771d38#artist"/>
   <mo:musicbrainz rdf:resource="&mba;d87e52c5-bb8d-4da8-b941-9f4928627dc8.html"/>
   <mo:wikipedia rdf:resource="http://en.wikipedia.org/wiki/ABBA"/>
   <owl:sameAs rdf:resource="http://dbpedia.org/resource/ABBA"/>
 </mo:MusicArtist>
</rdf:RDF>
```

Figure 2 Un fragment de code RDF au format RDF / XML décrivant le groupe ABBA

Autres sérialisations également **Turtle**, un format utilisé dans Figure 3 qui code RDF dans une langue triple et permet des raccourcis, comme « ; » pour répéter sujets ou «,» pour répéter sujet / prédicat paires.

```
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl:  <http://www.w3.org/2002/07/owl#> .
@prefix foaf:   <http://xmlns.com/foaf/0.1/> .
@prefix mo:   <http://purl.org/ontology/mo/> .

<http://www.bbc.co.uk/music/artists/d87e52c5-bb8d-4da8-b941-9f4928627dc8#artist>
    rdf:type mo:MusicArtist, mo:MusicGroup ;
    foaf:name "ABBA" ;
    foaf:homepage <http://www.abbasite.com/> ;
    mo:image <http://www.bbc.co.uk/music/images/artists/542x305/d87e52c5-bb8d-4da8-b941-9f4928627dc8.jpg> ;
    mo:member <http://www.bbc.co.uk/music/artists/042c35d3-0756-4804-b2c2-be57a683efa2#artist>,
              <http://www.bbc.co.uk/music/artists/2f031686-3f01-4f33-a4fc-fb3944532efa#artist>,
              <http://www.bbc.co.uk/music/artists/aebbb417-0d18-4fec-a2e2-ce9663d1fa7e#artist>,
              <http://www.bbc.co.uk/music/artists/ffb77292-9712-4d03-94aa-bdb1d4771d38#artist> ;
    mo:musicbrainz <http://musicbrainz.org/artist/d87e52c5-bb8d-4da8-b941-9f4928627dc8.html> ;
    mo:wikipedia <http://en.wikipedia.org/wiki/ABBA> ;
    owl:sameAs <http://dbpedia.org/resource/ABBA> .
```

Figure 3 Un fragment de code RDF au format Turtle décrivant le groupe ABBA

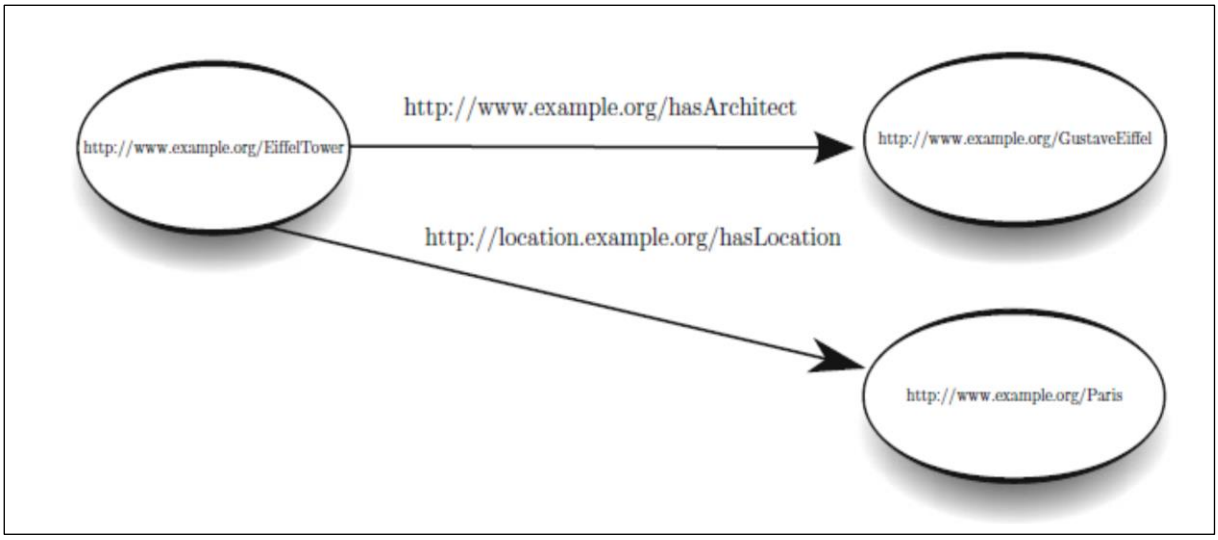

Figure 4 Fusion des triplets RDF

2.2.3. L'utilisation des espaces de noms

Prenons l'exemple de triplet suivant : ***ex : EiffelTower ex : hasArchitect ex : GustaveEiffel***. La ressource de départ dans le triplet est appelé **l'objet**, tandis que la liaison elle-même s'appelle le **prédicat** et la ressource d'arrivée dans le triplet est l'**objet**. Dans cet exemple **ex** représente l'abréviation de l'adresse *http://www.exemple.org/* , c'est ce qu'on appelle l'espace de nom.

En outre, le triplet semble similaire à des phrases en langage naturel simple, où le sujet et les objets sont des noms et le prédicat est un verbe. Du point de vue du web traditionnel, la principale caractéristique de RDF est que les liens dans RDF eux-mêmes ont un rôle URI requis. Par exemple, la relation entre Gustave Eiffel et de la Tour Eiffel pourrait être officialisée en tant que lien (comme indiqué dans la Figure 4). En RDF, les URI peuvent se référer à ces relations abstraites, même si ces URI peuvent ne pas être accessibles dans le même sens que les pages web. De cette manière, les prédicats RDF sont différents des liens hypertextes dans les systèmes traditionnels.

2.2.4. RDF et le principe de l'auto-description

En fournissant leur propre méthode de self description. Les triplets RDF peuvent être transportés d'un contexte à un autre. Dans un monde idéal où les conditions sont normales, comme lorsque l'URI dans le triplet peut être utilisé pour accéder à une page web pour décrire son contenu.

De même, une initiative appelée "**Linked Data**" tente de déployer des ensembles de données publics massifs, et son principe essentiel est de suivre le principe de l'auto Description [Bizer et al. 2007].

L'espoir est que le Web sémantique peut être considéré comme un réseau homogène de données liées, de sorte que l'agent peut découvrir **l'interprétation des données du Web sémantique simplement en suivant les liens.** Ces liens se rendront ensuite à d'autres données qui peuvent héberger des définitions formelles ou des descriptions en langage naturel informel et des représentations multimédias.

2.3. Les Technologies du web sémantique

Pour introduire la Sémantique au Web, le W3C a défini un certain nombre de langues, remplissant chacun un rôle particulier et la mise en œuvre d'une couche du Web sémantique [Luciano Floridi 2009]. Les différentes couches peuvent être trouvées dans la figure 5.

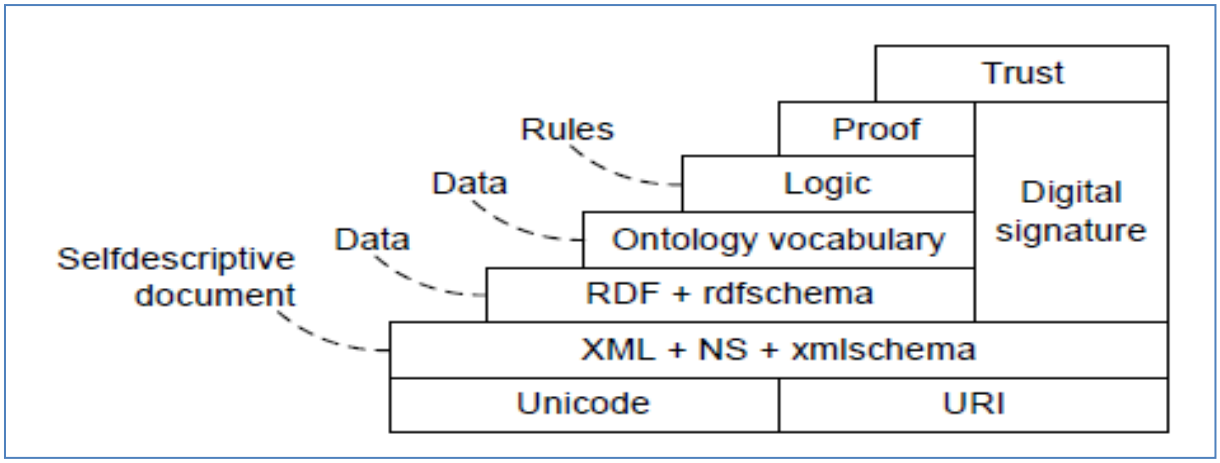

Figure 5 couches du sémantique web

Les couches de la partie inférieure de la pile, « *Unicode*, *URI*, *XML* + (*NS*) + *schéma XML*, *RDF* + *RDF Schéma*, et la couche de *vocabulaire de l'ontologie* (qui se compose de OWL) », sont largement en place. OWL est une recommandation du W3C depuis Février 2004. La recherche sur les couches supérieures restantes « *logique*, *Proof* and *Trust* », n'a pas encore abouti à des recommandations du W3C.

La couche Unicode garantit que toutes les langues aux dessus utilisent des jeux de caractères internationaux, tandis que la couche d'URI fournit des moyens pour identifier les ressources sur le Web sémantique. Tous les langages (y compris RDF (S) et OWL) adoptent XML (extensible Markup Language) comme syntaxe car ils sont construits au-dessus de la couche XML + NS + XML Schéma.

Le langage XML permet aux utilisateurs d'ajouter la structure dans leurs documents en utilisant un ensemble d'autodéfinition de tags, XML Schéma est utilisé pour définir la structure et le vocabulaire permis d'un document XML. Tel schéma en XML ne définit pas la sémantique des balises introduites. Il existe également de nombreux logiciels disponibles pour travailler et manipuler XML, ce qui facilite le travail des programmeurs qui utilise XML.

2.4. Le langage d'interrogation du Web sémantique - SPARQL

L'accès aux données est un élément important dans l'architecture du Web sémantique. Les principes de données liées permettent la publication et l'accès simples. Toutefois, ils ne prennent pas en charge les requêtes plus complexes parce que les données publiées en utilisant le paradigme de données liées ne doivent pas être accompagnées de mécanismes de requêtes plus sophistiquées. Prenons l'exemple à partir de MusicBrainz avec une requête pour obtenir des informations à propos d'ABBA. Maintenant, considérons une requête demandant des chanteurs d'ABBA qui étaient également membres d'autres groupes de musique, comme une telle situation n'est pas rare chez les musiciens. Même si un logiciel peut naviguer à partir de l'URI décrivant ABBA à chacun de ses membres, et plus tard à toutes les bandes où elle a été membre, cela peut être trop long pour certains cas d'utilisation. En outre, si une partie de l'URI ne pointent pas vers des adresses web réelles, il n'est même pas possible de trouver une réponse à une telle requête. [Maciej Janik & 2011]

Le langage de requêtes SPARQL pour RDF est conçu pour évaluer des requêtes sur des ensembles de données RDF et conçu pour traiter les requêtes d'une structure complexe, typiquement sur des données stockées dans des référentiels RDF. Le référentiel qui prend en charge SPARQL doit implémenter l'interrogation des données sous-jacentes en utilisant une syntaxe spécifique et un Protocol, avec RDF référentiels, l'accès à l'information est réalisé en posant des questions aux extrémités SPARQL.

SPARQL permet à un utilisateur de spécifier un URI (même ceux qui n'ont pas accessible sur le Web) et un motif graphique qui doit être comparée avec la base de connaissances avec des contraintes supplémentaires. Le modèle de graphe d'exemple pour poser des questions sur les différentes bandes que les musiciens d'ABBA chantaient est présenté dans Figure 6.

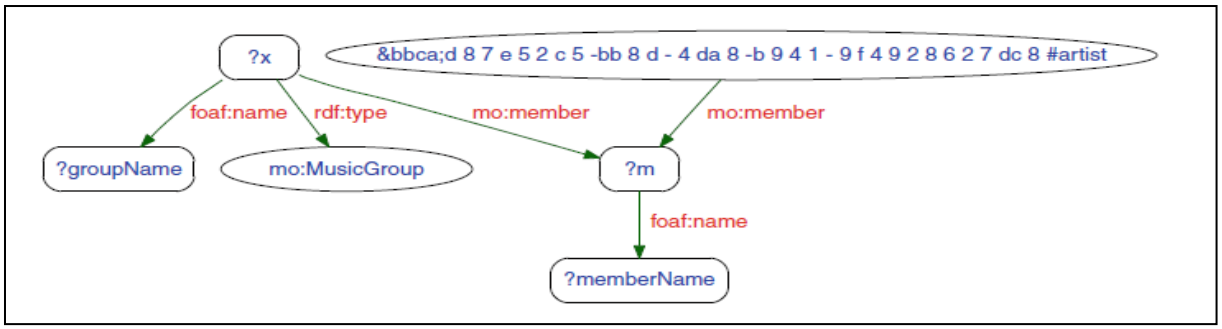

Figure 6 Représentation graphique de la clause WHERE de la requête pour des groupes de musique que les membres d'ABBA chantent

La requête peut être rédigée en SPARQL, tel que présenté dans la Figure 7. Une requête SPARQL est composée de sections qui définissent les différents aspects de la requête. ***Préfixe*** est utilisé pour abréger les URI, surtout pour la clarté et pour améliorer la lisibilité du modèle de graphe. Dans la section ***SELECT***, les utilisateurs peuvent spécifier l'information exacte qui les intéresse ; Alternativement, les utilisateurs peuvent demander des triplets comme résultat à une requête en utilisant la clause ***CONSTRUCT***.

```
PREFIX rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX mo:  <http://purl.org/ontology/mo/>
PREFIX foaf:    <http://xmlns.com/foaf/0.1/>

SELECT ?memberName ?groupName
WHERE  { <http://www.bbc.co.uk/music/artists/d87e52c5-bb8d-4da8-b941-9f4928627dc8#artist> mo:member ?m .
         ?x   mo:member ?m .
         ?x   rdf:type mo:MusicGroup .
         ?m   foaf:name ?memberName .
         ?x   foaf:name ?groupName }
FILTER (?groupName <>  "ABBA")
```

Figure 7 SPARQL pour les groupes de musique que les membres d'ABBA chantent

Le noyau d'une requête SPARQL est contenu dans la clause ***WHERE***. Ici, les utilisateurs définissent le modèle de graphe exact qui doit aller de pair avec les données du Web sémantique. Un modèle de graphe de base se compose de modèles individuels (sujet, prédicat, objet) qui sont reliés par des variables, formant un modèle qui sera comblé au cours du processus d'appariement. Dans l'exemple, les ressources d'assemblage comprennent le groupe inconnu et les membres d'ABBA. La clause WHERE est utilisée pour faire correspondre de la structure du graphe.

En option, la Clause WHERE est suivie d'une expression de filtre qui réduit les résultats retournés seulement pour les structures qui répondent à des critères spécifiques. L'exemple de requête filtre les noms des groupes de musique autres qu'ABBA.

2.5. Le Web Social Sémantique

2.5.1. Principe du web social sémantique

Le Web sémantique vise à fournir les outils qui sont nécessaires pour définir des normes extensibles et flexibles pour l'échange d'informations et l'interopérabilité. Un certain nombre de vocabulaires du Web sémantique ont atteint un déploiement à grande échelle : des exemples réussis comprennent **RSS** 1.0 pour la syndication d'informations (RSS 1.0 étant un format RDF, contrairement aux autres versions RSS), **FOAF** (Friend of a Friend) pour exprimer le profil personnel et information de réseautage social, et **SIOC** (communautés en ligne sémantiquement liés entre eux) pour les communautés d'interconnexion et les conversations distribués.

L'effort de Web sémantique est dans une position idéale pour faire des sites Web sociaux interopérables en fournissant des normes pour soutenir l'échange de données et d'interopérabilité entre les applications, permettant aux individus et aux communautés de participer à la création de l'information interopérable distribuée. La création d'un réseau social sémantiquement riche, réunissant des applications et des fonctionnalités du Web social avec des langages et des formats de représentation des connaissances du Web sémantique (Figure 8).

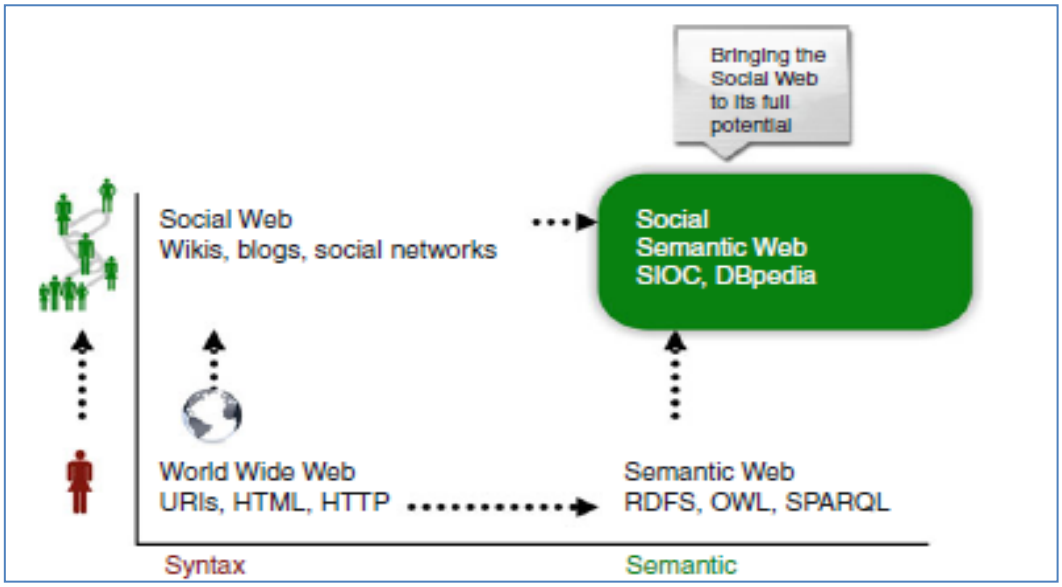

Figure 8 web social sémantique

Cette vision du Web se décompose de documents reliés entre eux, de données, et même des applications créés par les utilisateurs finaux eux-mêmes comme le résultat d'interactions sociales différentes. Ces documents et données sont écrits en utilisant des formats lisibles par la machine afin qu'ils puissent être utilisés à des fins que l'état actuel du Web Social ne peut pas atteindre sans difficulté.

Comme Tim Berners-Lee a déclaré dans un podcast 2005 [Tim Berners-Lee 2005], les technologies du web Sémantique peuvent soutenir les communautés en ligne même que les communautés en ligne supportent des données du Web sémantique en étant des sources pour relier les gens volontairement. Par conséquent, l'intégration entre les web Sémantique et le web social est double :

- D'une part, des efforts se concentrent sur l'utilisation des technologies du Web sémantique pour modéliser des données sociales. Avec des ontologies tels que **FOAF** et **SIOC**, les données sociales peuvent être représentées en utilisant des modèles communs et partagés, ce qui les rend donc plus facilement interopérables et transportables entre les applications.

- D'autre part, en s'appuyant sur le Web 2.0 on peut donner une longue avance vers la création d'une grande quantité de données du Web sémantique. En outre, les sites de réseautage Social Web peuvent contribuer à l'effort du Web sémantique. Les utilisateurs de ces sites offrent souvent des métadonnées sous la forme d'annotations et des étiquettes sur les photos, notes, liens de blog, etc. Les utilisateurs du site sociaux créent déjà un vocabulaire étendu et annotations sémantiquement riches à travers ***folksonomie***. De cette façon, les réseaux sociaux et sémantiques peuvent se compléter mutuellement.

On ne peut pas créer des applications web sémantique utiles sans avoir des données pour les alimenter, et vous ne pouvez pas produire des données sémantiquement riches sans les applications intéressantes eux-mêmes. Depuis le Social Web contient un tel contenu sémantiquement riche, des applications intéressantes alimentées par les technologies du Web sémantique peuvent être créés immédiatement. Le Web sémantique sociale offre un certain nombre de possibilités en termes d'accroissement de l'automatisation et de la diffusion de l'information qui ne sont pas facilement réalisables avec les applications actuelles de logiciels sociaux.

2.5.2. Les ontologies pour le Web social sémantique

L'ontologie fournit un modèle commun pour représenter des informations sémantiquement riche sur le Web Sémantique. En outre, en utilisant des langages de représentation standard, tels que RDF (S) / OWL, ces ontologies peuvent être partagées entre les services, de sorte que les données deviennent interopérables entre les applications distribuées. Dans le domaine du web social sémantique [John G. Breslin & ALL 2011], les ontologies peuvent être ensuite utilisées pour représenter de manière uniforme les différents produits et partagé dans des sites Web sociaux : les communautés, les gens, documents, étiquettes, etc. Dans cette section, certaines des ontologies les plus populaires pour le Web social sont décrites.

A. *FOAF – Friend of a Friend*

Friend -of-a-Friend (FOAF) projet [1] a été lancé par Dan Brickley et Libby Miller en 2000, elle définit un vocabulaire couramment utilisé pour décrire les personnes et les relations entre eux.

FOAF est un *RDF vocubulary* pour décrire les gens, leurs relations et leurs propriétés. Un profil FOAF décrit une personne, quelques faits de cette personne (par exemple, les intérêts, les projets en cours), et des connexions à d'autres personnes (personnes « connus » par cette personne) d'une manière lisible par la machine et défini sémantiquement.

Il permet aux gens de créer des pages Web lisibles par la machine pour les personnes, les groupes, les organisations et autres concepts connexes.

Les principales classes dans le vocabulaire FOAF (Figure 9, comme illustré par Dan Brickley.) Comprennent **Foaf : Person** (pour décrire les personnes), **foaf : OnlineAccount** (pour le détail des comptes d'utilisateurs en ligne qu'ils détiennent), et **foaf : Document** (pour les documents que les gens créent). Certaines des propriétés les plus importantes sont **foaf : knows** (utilisé pour créer un lien de connaissance), et **FOAF : topic_interest** (utilisé pour pointer vers des ressources qui représentent l'intérêt qu'une personne peut avoir).

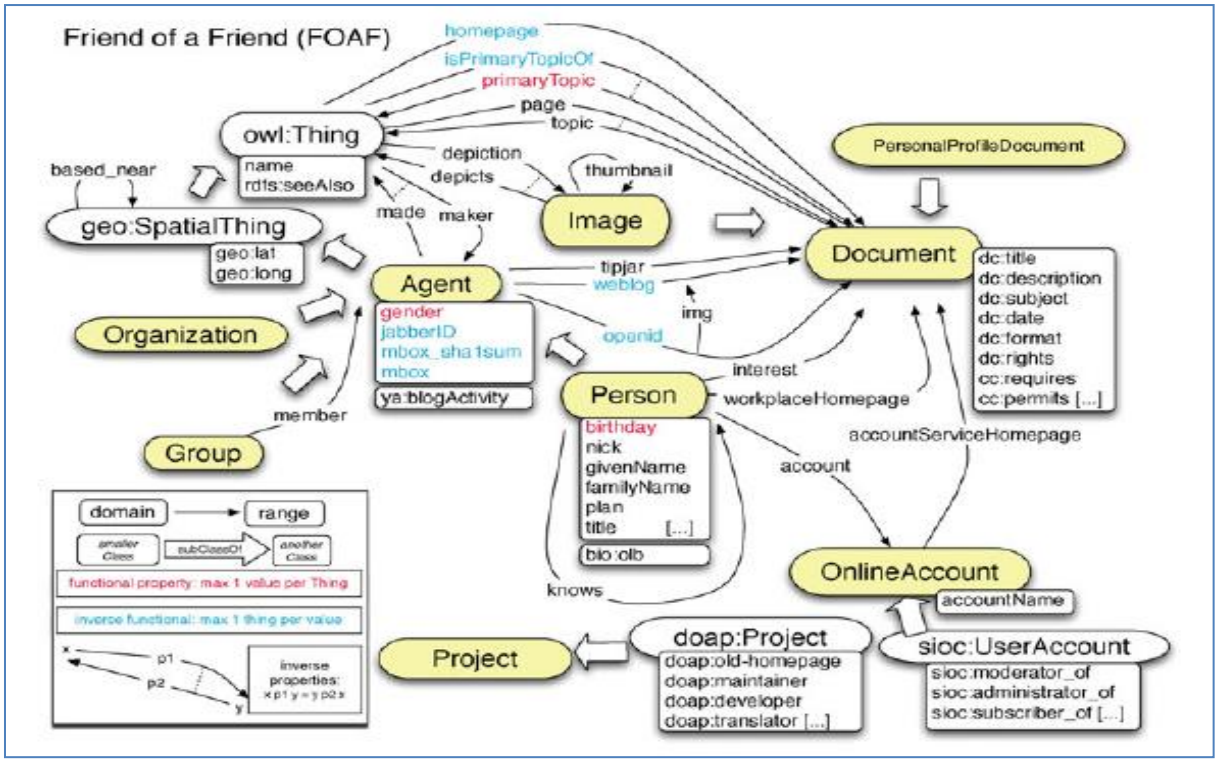

Figure 9 Conditions friend-of-a-Friend: Mise à jour de la photo de Dan Brickley CC (http://www/flickr.com/ photos/danbri/1855393361 /)

[1] http://www.foaf-project.org

B. SKOS - Simple Knowledge Organization System

FOAF peut être intégré avec d'autres vocabulaires du Web Sémantique, tels que SIOC et **SKOS** - Simple Knowledge Organization System.
Certains services de réseautage social de premier plan qui exposent des données à l'aide de FOAF comprennent **Hi5** (un site de réseautage social), LiveJournal (un réseau social et site de la communauté des blogueurs), Identi.ca (un site de Microblogging).
La représentation de la connaissance d'une personne et leurs amis seraient atteints par un fragment FOAF similaire à celle de la Figure 10.

```
@prefix foaf: <http://xmlns.com/foaf/0.1.>.
<http://www.johnbreslin.com/foaf/foaf.rdf#me> a foaf:Person;
   foaf:name "John Breslin";
   foaf:mbox <mailto:john.breslin@deri.org>;
   foaf:homepage <http://www.johnbreslin.com/>;
   foaf:nick "Cloud";
   foaf:depiction <http://www.johnbreslin.com/images/foaf_photo.jpg>;
   foaf:topic_interest <http://dbpedia.org/resource/SIOC>;
   foaf:knows [
     a foaf:Person;
     foaf:name "Sheila Kinsella";
     foaf:mbox <mailto:sheila.kinsella@deri.org>
   ];
   foaf:knows [
     a foaf:Person;
     foaf:name "Smitashree Choudhury";
     foaf:mbox <mailto:smitashree.choudhury@deri.org> ].
```

Figure 10 Représentation de la connaissance d'une personne et leurs amis en FOAF

C. Semantically Interlinked Online Communities SIOC

SIOC (Semantically Interlinked Online Communities) est un vocabulaire permettant de décrire des objets couramment utilisés sur les sites communautaires et leurs relations. Il est défini en utilisant RDF. SIOC est une application du web sémantique pour décrire des blogs, des forums, des wikis,...etc. En plus d'une ontologie, le projet fournit également différents outils pour utiliser le vocabulaire. Il utilise des objets définis dans d'autres ontologies, comme FOAF (pour décrire les personnes impliquées), SKOS, et RSS (pour décrire les contenus) [sioc-project].

SIOC vise à interconnecter le contenu de la communauté en ligne liée à partir de plates-formes telles que les blogs, les forums et autres sites sociaux, en proposant une ontologie légère pour décrire la structure de ses activités dans les communautés en ligne, en combinaison avec le vocabulaire FOAF pour décrire les personnes et leurs amis, et le modèle SKOS pour l'organisation des connaissances.

SIOC évolue pour décrire non seulement les plates-formes de discussion classiques, mais aussi de nouveaux mécanismes de partage de contenu et de communication sur le Web. À l'heure actuelle, une grande partie du contenu en cours de création sur des sites sociaux (manifestations, des signets, des vidéos, etc.) est commentée et annotée par d'autres. Les sites sociaux déconnectés nécessitent une ontologie pour l'interopérabilité, et en raison du fait qu'il y a beaucoup de données sociales avec une sémantique inhérente contenue dans ces sites, il existe un potentiel d'impact important dans le déploiement réussi d'une ontologie SIOC.

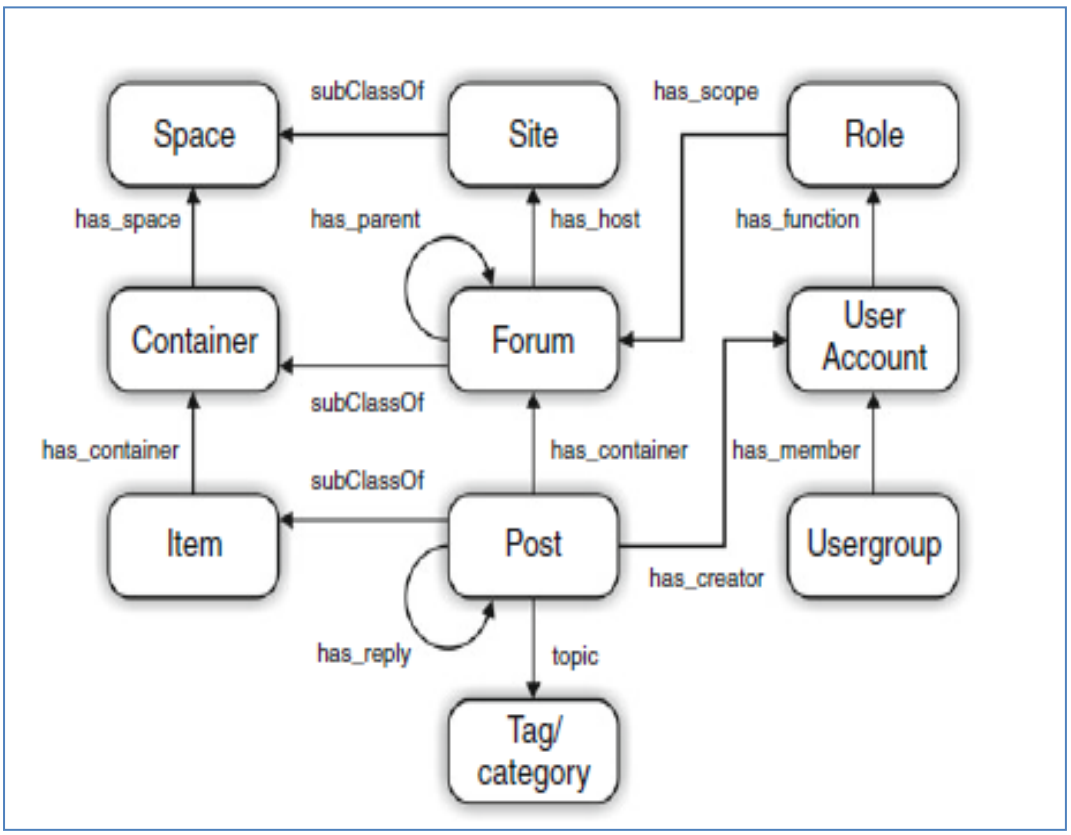

Figure 11 L'ontologie SIOC

2.6. Conclusion

Dans ce chapitre nous avons donnés une vision globale du web sémantique ainsi que son langage de représentation et de description des connaissances RDF permettant de donner de la sémantique aux documents dans le but d'être compris et exploités par la machine. Le web social a profité des technologies et des langages du web sémantique pour faire apparaitre un nouveau web social plus riche appelé web social sémantique.

Le web social sémantique se sert des ontologies telles que FOAF, SKOS, …, etc. pour créer des profils utilisateurs plus étendus pouvant être exploités par d'autres applications.

Dans le chapitre qui suit nous présenterons les données liées et l'émergence de ces derniers dans le web et l'appréciation du web de données.

Chapitre 3
Linked Data

3.1. Introduction

Tout comme le World Wide Web a révolutionné la façon de trouver des documents et de s'en servir, il peut révolutionner la façon de *découvrir* les données, d'y *accéder*, de les *intégrer* et de les *utiliser*. Le Web est le media idéal pour ces processus, en raison de son ubiquité — sa nature distribuée permettant les changements d'échelle –, ainsi que de la maturité de sa technologie. Tim Berners-Lee, directeur du W3C, a inventé et défini le terme Linked Data et son synonyme Web of Data au sein d'un ouvrage portant sur l'avenir du Web sémantique.

Ce chapitre s'intéresse à la façon dont un ensemble de principes et de technologies, connu sous le nom de *données liées* (*Linked Data*), exploite la philosophie et l'infrastructure du Web pour permettre le partage de données et leur réutilisation à grande échelle.

3.2. La justification des données liées

Afin de comprendre le concept et la valeur des données liées, il est important d'appréhender les mécanismes actuels d'échange et de réutilisation de données sur le Web.

- **La structuration permet des traitements sophistiqués :**

Un facteur-clé dans la réutilisation des données est la façon dont elles sont structurées. Plus cette structure est régulière et bien définie, plus cela facilite la création d'outils pour traiter et réutiliser les données de manière fiable. [Tom Heath and Christian Bizer 2011]

Le langage le plus utilisé pour créer des sites web c'est le HTML, qui se concentre sur la présentation textuelle de documents plutôt que de la structuration des données. Ces derniers s'entremêlant au texte, ce qui rend difficile pour des applications logicielles d'extraire des bribes structurées à partir de pages HTML.

Pour résoudre ce problème, de nombreux *micro formats* ont été inventés. Ils peuvent être utilisés pour publier des données structurées décrivant des types spécifiques d'entités, comme des personnes, des organisations, des événements, des critiques et des notes. Ils servent à incorporer de manière très précise des données dans des pages HTML, ce qui permet aux applications de les extraire sans ambiguïté.

Mais, les micros formats, Ils ne fournissent qu'un petit ensemble d'attributs pour les décrire, il est impossible d'exprimer les relations entre les entités. Par conséquent, les micros formats ne sont pas appropriés pour le partage de données sur le Web.

Les API web ont une méthode plus générique pour rendre les données structurées disponibles sur le Web. Elles fournissent, par le biais de requêtes, un accès simple a des données structurées via le protocole HTTP : des exemples connus sont l'API d'Amazon, *Product Advertising*, ou celle de *Flickr*.

L'avènement de cette méthode a conduit à une explosion de petites applications composites qui combinent les données de plusieurs sources, chacune d'elles étant accessible via une API propre au fournisseur de données. Alors que les avantages d'un accès à des données structurées par du code sont indiscutables, l'existence d'une API pour chaque jeu de données crée une situation dans laquelle l'intégration des données de chaque source dans une application exige des efforts importants. Les programmeurs doivent connaitre les méthodes pour récupérer les données de chaque API et écrire du code spécifique pour accéder aux données de chaque source.

- **Les hyperliens connectent les données distribuées :**

Linked Data fournit une solution technique pour faciliter la découverte des données dans le web par les *agents web*, tels que les navigateurs et les robots des moteurs de recherche, Et donc la facilité de découverte peut être appliquée à des données sur le Web.

3.3. Définition de Linked Data

Le Web sémantique n'est pas seulement de mettre les données sur le web. Il s'agit de faire des liens, de sorte qu'une personne ou une machine peuvent explorer le web de données d'une manière simple et efficace. Avec les données liées, à partir d'un lien on peut accéder a d'autre données connexes.

Comme le web de l'hypertexte, le web de données est construit avec des documents sur le web. Cependant, contrairement au web de l'hypertexte, pour les données liées les liens sont décrits par RDF. L'URI identifier n'importe quel objet ou un concept [Tim Berners-Lee 2006]. Les quatres régles de linked data définit par Tim Berners-Lee sont :

- Utilisez URI comme noms pour les choses.
- Utiliser HTTP URIs de sorte que les gens peuvent regarder ces noms.
- Quand quelqu'un regarde un URI, fournir des informations utiles, en utilisant les standards (RDF, SPARQL).
- Inclure des liens vers d'autres URI. afin qu'ils puissent découvrir plus de choses.

3.4. Le Cycle de vie de Linked Data

Les différentes étapes du cycle de vie des données liées sont illustrées dans la Figure 12.

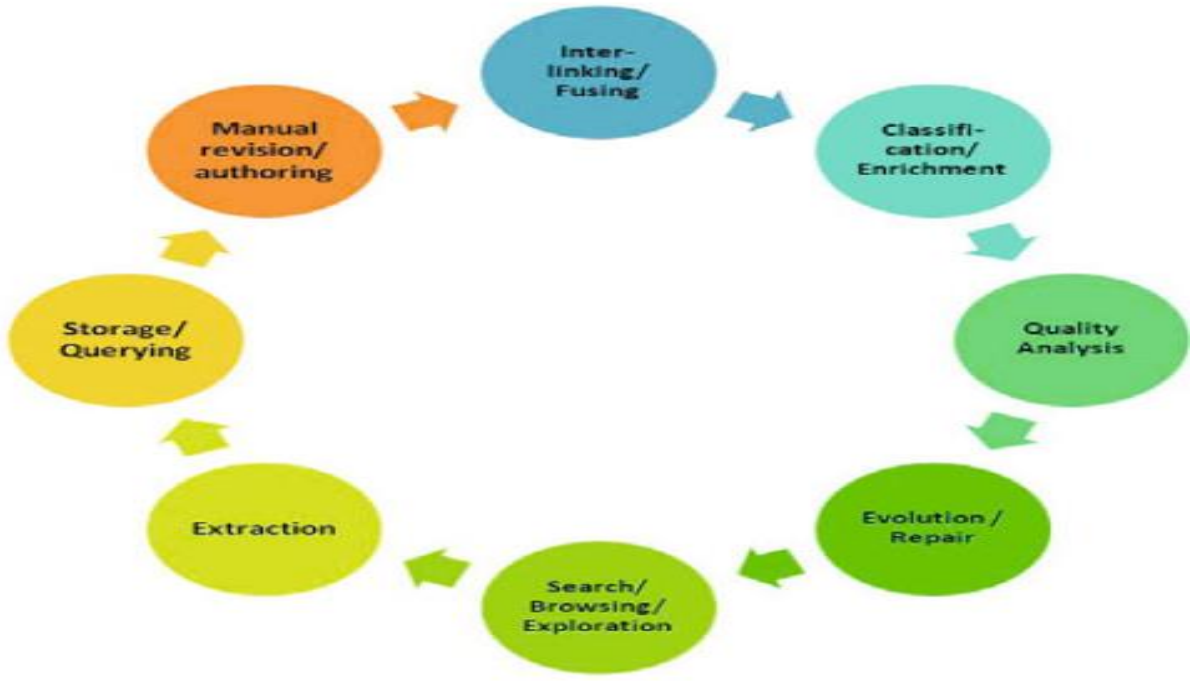

Figure 12 cycle de vie de Linked data

Les informations représentées sous forme structurée ou sous d'autres formalismes de représentation structurés ou semi-structurés doivent être mises en correspondance avec le modèle de données RDF **(Extraction)** [Jens Lehmann & all 2011].

Une fois qu'il y a une masse critique de **données RDF**, **des mécanismes** doivent être mis en place pour **stocker**, **d'indexer** et de **rechercher** efficacement ces données RDF **(Storage & Interrogation)**.

Les utilisateurs doivent avoir la possibilité de créer de nouvelles informations structurées ou de corriger et développer les partenariats existants **(Authoring).**

Si différents éditeurs de données fournissent des informations sur les mêmes entités associées, les liens entre ces différents éléments d'information doivent être mis en place **(linking)**.

Depuis Linked Data comprend principalement les données d'exemple, nous observons un manque de classification, de la structure et le schéma d'informations. Cette carence peut être abordé par des approches d'enrichissement de données avec les structures de plus haut niveau afin d' être en mesure de regrouper et interroger les données de manière plus efficace (**enrichissement**) .

Comme avec le Document Web, les données Web contient une variété d'informations de qualité différente. Par conséquent, il est important de concevoir des stratégies pour évaluer la qualité des données publiées sur le Web de données **(Qualité Analyse).**

Une fois que les problèmes sont détectés, les stratégies pour la réparation de ces problèmes et de soutenir l'évolution des données liées sont nécessaires **(Evolution et réparation)**.

Dernier point mais non moins, les utilisateurs doivent être habilités à **parcourir**, **rechercher** et **explorer** les informations de structure disponible sur le Web de données de manière rapide et convivial **(Recherche, navigation et Exploration)**.

3.5. Les principes de données liées

Le terme « **donnés liées »** se réfère à un ensemble de bonnes pratiques à mettre en œuvre pour **publier** et **lier** des données structurées **sur le Web**. Ces pratiques ont été introduites par Tim Berners-Lee dans *Linked Data* [sioc-project]. Ses notes sur l'architecture du Web sont connues en tant que *principes des données liées.* Ces principes sont résumés sur les points suivants:

- Nommer les éléments avec des URI ;
- Utiliser des URI HTTP, pour que l'on puisse rechercher/consulter ces noms ;
- Fournir des informations nécessaires sous forme de standards (RDF, SPARQL) lors d'une recherche d'URI ;
- Inclure des liens vers d'autres URI qui permettent de découvrir d'autres éléments.

3.5.1. Nommer les éléments avec des URI

Le premier principe des données liées recommande d'utiliser des références URI pour identifier non seulement des documents et du contenu digital, mais aussi des objets réels et des concepts abstraits. Cela peut être vu comme une extension des principes du Web pour comprendre tout objet ou concept.

Les techniques de réalisation de ce principe :

Pour publier des données sur le Web, il faut d'abord identifier les éléments du domaine d'intérêt. Il s'agit des éléments dont les propriétés et les relations seront décrites dans les données ; il peut s'agir de documents web, d'entités réelles et de concepts abstraits. Puisque les données liées s'appuient directement sur l'architecture du Web [R. Moats 2007], le terme *ressource* est utilisé pour nommer ces éléments dignes d'intérêt, qui sont, à leur tour, identifiés par des URI HTTP.

La Figure 13 montre l'utilisation d'URI HTTP pour identifier des entités réelles et leurs relations. Sur cette photo de l'équipe de tournage de *Big Lynx* au travail, on voit le cameraman principal, Matt Briggs, avec ses deux assistants, Linda Meyer et Scott Miller, identifiés par des URI HTTP de l'espace de noms *Big Lynx*. La relation (ils se connaissent) est représentée par des lignes, avec une URI de type *http://xmlns.com/foaf/0.1/knows*.

Figure 13 Les URI sont utilisées pour identifier des gens et les relations qui les joignent.

Ces données liées n'utilisent donc que des URI HTTP et cela pour deux raisons :

- Les URI HTTP fournissent une manière simple de créer des noms globalement uniques, de façon décentralisée, puisque n'importe qui possédant un nom de domaine peut créer ou déléguer la création de références URI.

- Elles servent de nom mais aussi de moyen d'accès à l'information décrivant l'entité identifiée.

3.5.2. Utiliser des URI HTTP, pour accéder ou rechercher des éléments

Le protocole HTTP est le mécanisme universel d'accès au Web. Dans le Web classique, des URI HTTP combinent l'identification unique et le mécanisme de récupération simple. Ainsi, le deuxième principe recommande l'utilisation de ces dernières pour identifier des objets et des concepts abstraits, afin que ces URI soient déréférencées (autrement dit, que l'on puisse récupérer le contenu pointé) par le protocole HTTP et traduites en une description de l'objet identifié ou du concept.

Rendre URI différenciables:

Toute URI HTTP doit être différenciables, ce qui signifie que les clients HTTP peuvent rechercher l'URI en utilisant le protocole HTTP et récupérer une description de la ressource identifiée par l'URI.

L'URI identifie les objets du monde réel, il est essentiel de ne pas confondre les objets eux-mêmes avec les documents Web qui les décrivent. La pratique courante d'utiliser différentes URI pour identifier l'objet dans le monde réel et le document qui le décrit, pour être sans ambiguïté.

Cette pratique permet la déclaration séparée pour être faites sur un objet et sur un document qui décrit cet objet. Par exemple, la date de création d'une personne peut être différente de la date de création d'un document qui décrit cette personne.

Etre capable de distinguer les deux déclarations grâce à l'utilisation de différentes URI est essentielle à la cohérence du Web de données. [R. Moats 2007] .

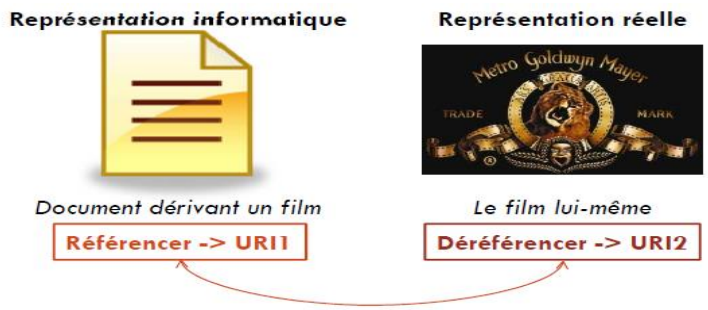

Figure 14 Séparation des déclarations de l'objet et du document qui le décrit

L'idée sous-jacente suppose que les clients HTTP envoient des-en-têtes HTTP avec chaque requête pour indiquer les types de documents qu'ils préfèrent. Les serveurs inspectent ces-en-têtes et répondent de façon appropriée : si l'en-tête indique que le client préfère le HTML, le serveur lui envoie un document HTML et, s'il préfère le RDF, le serveur lui envoie un document RDF.

Il existe deux stratégies pour créer des URI qui identifient des objets réels. Toutes deux s'assurent que les objets et les documents qui les décrivent ne sont pas confondus et que les humains autant que les machines peuvent récupérer des représentations appropriées. Ces stratégies sont appelées *URI 303* et *URI avec ancre*.

3.5.3. Utilisation de RDF et négociation de contenu

Pour que d'avantage d'applications différentes accèdent au contenu web, il est important d'utiliser un format de contenu standardisé. Le choix de HTML comme format de documents dominant a été un facteur prépondérant dans la croissance du Web.

Le troisième principe conseille donc un modèle de données simple fondé sur une structure en graphe conçue spécifiquement pour le contexte du Web. Les navigateurs HTML affichent généralement les représentations RDF sous forme de code RDF brut, ou simplement les télécharger sous forme de fichiers RDF sans les afficher, ce qui n'est pas très utile pour l'utilisateur moyen.

Pour résoudre ce problème d'affichage on peut utiliser un mécanisme [Tim Berners-Lee 1999] HTTP appelé négociation de contenu. Les clients HTTP envoient les en-têtes HTTP avec chaque requête pour indiquer quels types de représentation qu'ils préfèrent. Les serveurs peuvent inspecter ces en-têtes et sélectionnez une réponse appropriée.

Si les en-têtes indiquent que le client préfère HTML, le serveur peut générer une représentation HTML. Si le client préfère RDF, le serveur peut générer RDF. [Chris Bizer 2007]

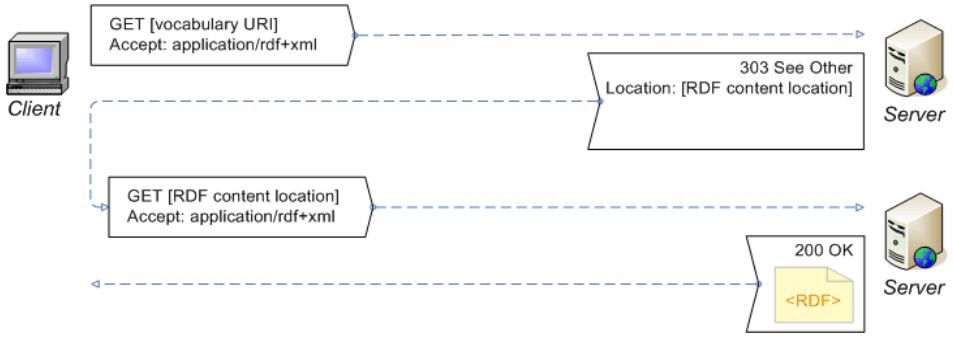

Figure 15 Négociation de contenu entre le client et le serveur

3.5.4. Inclusion des liens externes

Le quatrième principe propose d'utiliser les hyperliens afin de connecter toutes sortes d'éléments et pas seulement des documents web : ainsi, un **hyperlien** peut être **défini entre** une **personne** et un **lieu** ou entre un **lieu** et une **entreprise**.

Par exemple : un hyperlien de type **« *ami de* »** pourrait être défini entre deux personnes, un autre type de lien « ***habite près de »*** entre une personne et un endroit. Pour les distinguer, **ces hyperliens sont appelés liens RDF.**

Le principe de Linked Data est de mettre des liens RDF pointant vers d'autres sources de données sur le Web. Ces *liens RDF externes* sont fondamentaux pour le Web de données car elles sont le ciment qui relie les îles de données dans un espace de données interconnectées et mondial, car elles permettent aux applications de découvrir les sources de données supplémentaires dans un mode *de suivi de données.*

Techniquement, **un lien RDF externe** est un **triplet RDF,** où l'objet du triplet est une référence URI dans l'espace d'un ensemble de données, tandis que le prédicat et /ou l'objet du triplet sont des références d'URI pointant dans les espaces de noms d'autres ensembles de données.

Déréférencer ces URI donne une description de la ressource liée fourni par le serveur distant. Cette description contient généralement des liens RDF supplémentaires qui pointent vers d'autres URI qui à leur tour, peuvent également être déréférencé, et ainsi de suite. C'est ainsi que des descriptions de ressources individuelles sont tissées dans le Web de données. C'est aussi la façon dont le Web de données peut être consulté en utilisant un navigateur de données liées ou analysé par le robot d'un moteur de recherche. [Tim Berners-Lee 2006].

3.6. Le Web des données

Un nombre important d'individus et d'organisations ont adopté des Linked Data comme un moyen de publier leurs données. Le résultat est un espace de données globale que nous appelons le *Web de données*. Le Web des données constitue un graphique géant mondial composée de milliards de déclarations RDF à partir de nombreuses sources qui couvrent toutes sortes de sujets, tels que les emplacements géographiques, des personnes, des entreprises, des livres, des publications scientifiques, des films, de la musique, de la télévision et de programmes radiophoniques, les gènes, les protéines, les médicaments et les essais cliniques, des données statistiques, des résultats de recensement, les communautés en ligne et des examens.

Les entités sont reliées par des liens RDF, qui permet la création d'un graphique de données globale qui se ramifie avec des sources de données et permet la découverte de nouvelles sources de données. Cela signifie que les applications n'ont pas un ensemble fixe de sources de données, mais ils peuvent découvrir de nouvelles sources de données au moment de l'exécution et cela en suivant les liens RDF.

3.6.1. Démarrage du Web de données

Les origines de ce Web des données se trouvent dans les efforts de la communauté de recherche du Web sémantique et en particulier dans les activités du W3C *Linked Open Data (LOD),* fondée en Janvier 2007. L'objectif fondateur du projet, qui a donné naissance à une communauté Linked Data dynamique et en expansion, était d'amorcer le Web de données en identifiant l'ensemble des données existantes disponibles sous les licences ouvertes, les convertir en RDF selon les principes de données liées, et de les publier sur le Web. Comme une question de principe, le projet a toujours été ouvert à toute personne qui publie des données selon les principes de Linked Data. Cette ouverture est un facteur probable dans la réussite du projet en amorçant le Web de données.

Les Figures 16, 17 montrent comment le nombre d'ensembles de données publiées sur le Web comme Linked Data a augmenté depuis la création du projet Open Data. Chaque nœud dans le diagramme représente un ensemble publié comme Linked Data de données distinctes. Les arcs indiquent l'existence de liens entre les éléments des deux ensembles de données. Les arcs lourds correspondent à un plus grand nombre de liaisons, alors que des arcs bidirectionnels indiquent que les liaisons vers l'extérieur existent entre les deux ensembles de données.

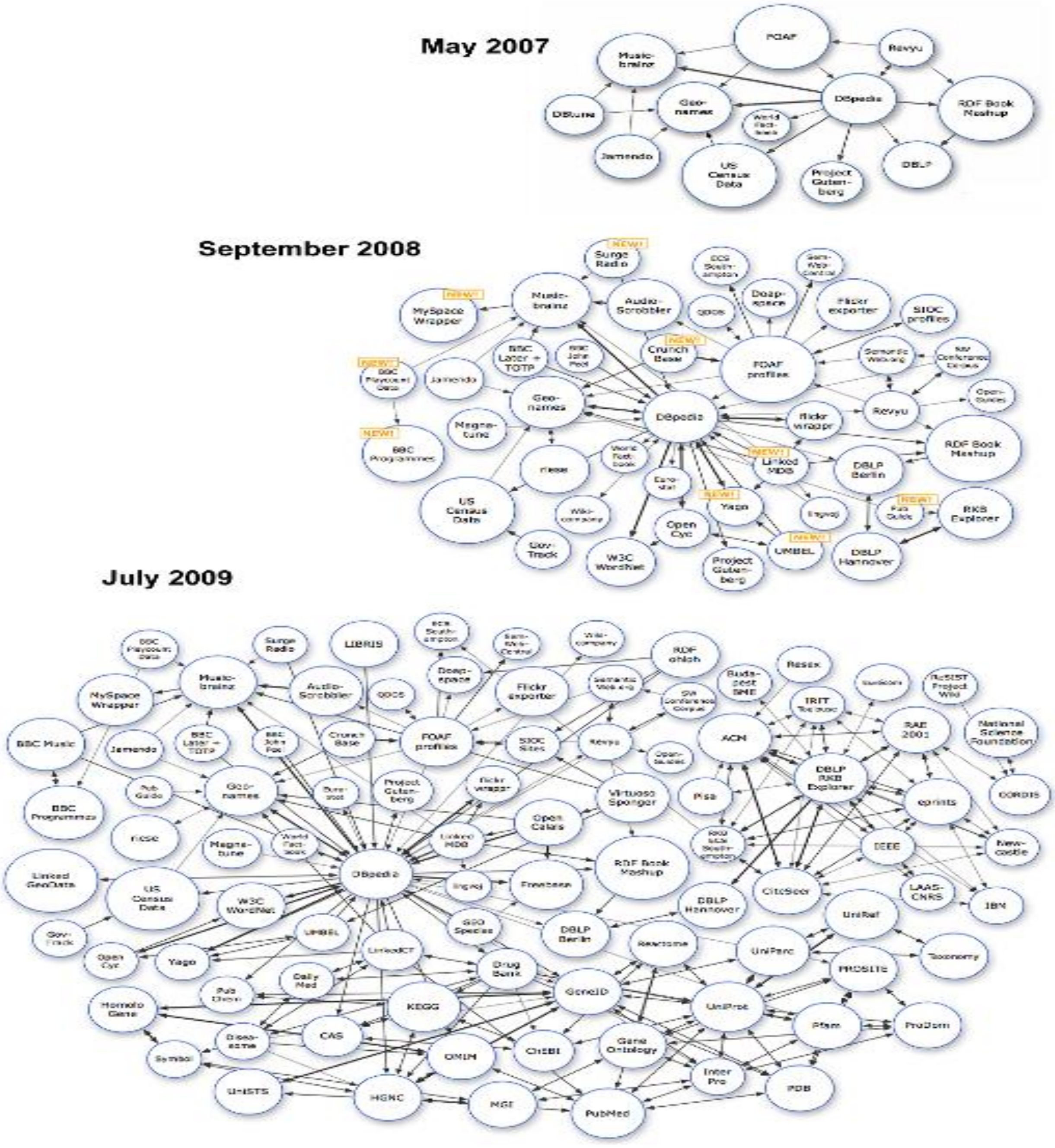

Figure 16 Croissance du nombre d'ensembles de données publiées sur le Web comme Linked Data.

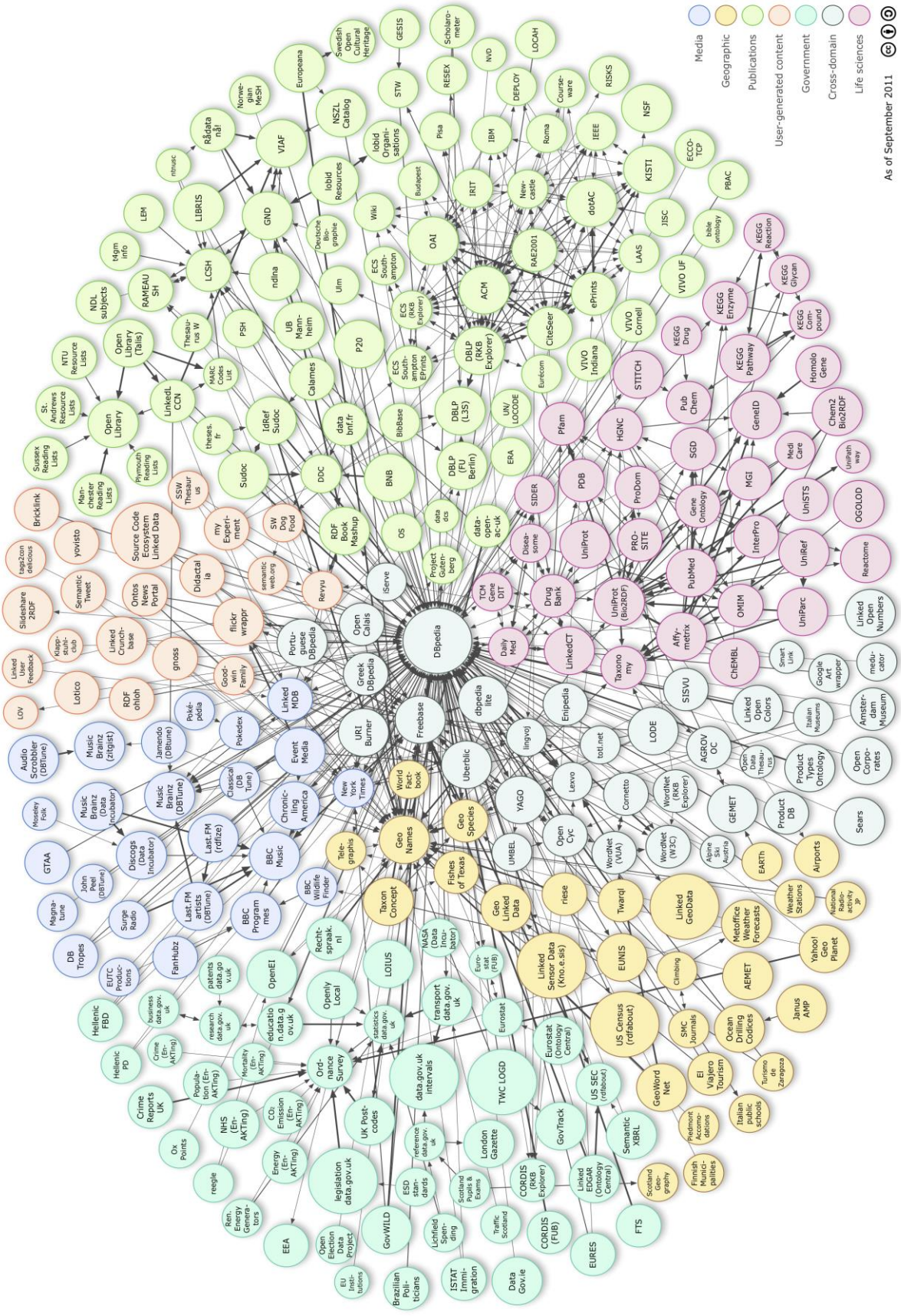

Figure 17 Linked Open Cloud de données à partir de septembre 2011. Les couleurs classent des ensembles de données par domaine d'actualité.

3.6.2. Topologie du Web de données

Cette section propose un aperçu de la topologie du Web de données telle qu'elle était en septembre 2011. Les jeux sont classés selon les domaines suivants : géographie, gouvernement, médias, bibliothèques, sciences de la vie, commerce, contenu généré par les utilisateurs, jeu de données à la croisée de plusieurs domaines.

- **Données multi domaines**

Parmi les premiers jeux de données, certains n'étaient pas propres à un domaine en particulier mais en couvraient plusieurs. L'exemple principal de données liées multi domaines est ***DBpedia***, un jeu de données automatiquement extraites des pages publiques de Wikipédia. Une URI DBpedia est automatiquement attribuée à l'article Wikipédia correspondant.

- **Données provenant des médias :** l'une des premières organisations à avoir reconnu le potentiel des données liées et à avoir adopté ses principes et technologies dans sa stratégie de publication et d'organisation de contenu est la BBC. Dans ce cas, le but de l'utilisation de RDF n'était pas d'exposer des données liées pour une consommation par des tiers, mais de faciliter la gestion et le stockage en interne des données ainsi que leur intégration dans un domaine

- **Bibliothèques et éducation :** l'obligation pour les bibliothèques de fournir de nouveaux moyens de découverte et leur grande expérience dans la production de données structurées de qualité font de ces établissements des acteurs complémentaires naturels des données liées. Ce domaine a vécu plusieurs développements récents dans l'indexage des catalogues

- **Données des sciences de la vie :** le projet des données liées a été largement adopté par la communauté des sciences de la vie en tant que technologie pour connecter divers jeux de données utilisées par les chercheurs.

- **Vente et commerce :** le livre RDF Book Mashup fournit un exemple récent de publication de données liées sur le commerce et la vente au détail.

- **Contenu généré par les utilisateurs et les médias sociaux :** certains des jeux de données les plus récents sont fondés sur la conversion ou des surcouches de sites Web 2.0 contenant de larges volumes de contenus générés par les utilisateurs. Cela a produit des jeux de données et des services tels que *DBpedia* et *FlickrWrappr*. Ils ont été complétés par des sites de contenus produits par les utilisateurs, sites bâtis avec des supports natifs pour les données liées, par exemple Revyu.com pour les systèmes de commentaires et de classements par votes et *Faviki* pour annoter du contenu de pages web avec des URI de données liées.

3.6.3. Relier les données du web

Pour relier deux jeux de données qui sont décrits chacun par une ontologie, il est nécessaire d'appliquer une méthode de comparaison des ressources. Le résultat de la méthode de comparaison, automatique ou manuelle est un ensemble de relations owl:sameAs entre ces ressources. [Jérôme Euzenat& ALL 2011]

a. Alignement manuel des ressources :

Dans ce cas les ressources sont alignées par une observation manuelle. On peut faire appel à des outils collaboratifs si on a un grand jeu de données.

b. Mise en correspondance des identifiants

Dans ce cas une simple transformation est effectuée pour mettre en correspondance les identifiants des ressources où un ensemble de règles est défini pour identifier les ressources équivalentes à partir de leur identifiant.

c. Alignement de données avec ontologie commune

Le rôle du système d'alignement des données décrites avec la même ontologie est de rechercher les ressources de même type afin de détecter celles qui sont équivalentes. Pour cela le système va comparer les propriétés des ressources et construire une mesure de similarité.

d. Ontologies différentes et alignement implicite

Si les données à relier sont décrites par des ontologies hétérogènes un alignement entre les ontologies doit être effectué pour indiquer au système d'alignement les correspondances entre les entités des ontologies. Le système fonctionne ensuite de façon similaire à un système à une seule ontologie. Dans ce cas, l'alignement est spécifié de manière implicite par l'utilisateur.

e. Ontologies différentes et alignement explicite

La différence entre l'alignement explicite et implicite réside dans l'utilisation d'un outil pour faire l'alignement entre les deux ontologies. L'utilisateur n'a pas dans ce cas à décrire d'alignement.

3.6.4. Méthodes pour publier les données comme Linked Data

Les données doivent répondre aux exigences minimales suivantes :

- être considérées comme « publiés sous forme de données liées sur le Web » [Chris Bizer 2007]

- Les choses doivent être identifiées avec HTTP URIs. Si une telle URI est déréférencée demandant le type MIME application / RDF + XML, une source de données doit retourner une description RDF / XML de la ressource identifiée.

- URI qui identifient les ressources non-information doit être mise en place dans l'une de ces deux façons : soit la source de données doit renvoyer une réponse HTTP contenant une *303 redirection HTTP* vers une ressource d'information décrivant la ressource non-information, ou bien l'URI de la ressource non-information doit être formée en prenant l'URI de la ressource d'information connexe et l'ajout d'un *identifiant de fragment*.

- Les descriptions RDF devraient également contenir des liens RDF aux ressources fournies par d'autres sources de données, de sorte que les clients peuvent naviguer sur le Web de données dans son ensemble en suivant les liens RDF.

- **Utilisation des fichiers RDF statiques**

La façon la plus simple de publier les données liées est de produire des fichiers RDF statiques, et les télécharger sur un serveur web. Cette approche est généralement choisie dans les situations où les fichiers RDF sont créés manuellement, par exemple lors de la publication des fichiers personnels FOAF ou des vocabulaires RDF.

- **Utilisation des bases de données relationnelles**

Si les données sont stockées dans une base de données relationnelle, il est recommandé d'utiliser des outils pour publier cette dernière. Parmi ces outils on distingue le *D2R Server*. Ce serveur s'appuie sur une cartographie pour faire le mapping entre les schémas de la base et les termes RDF cibles et fournit une extrémité SPARQL pour interroger la base de données.

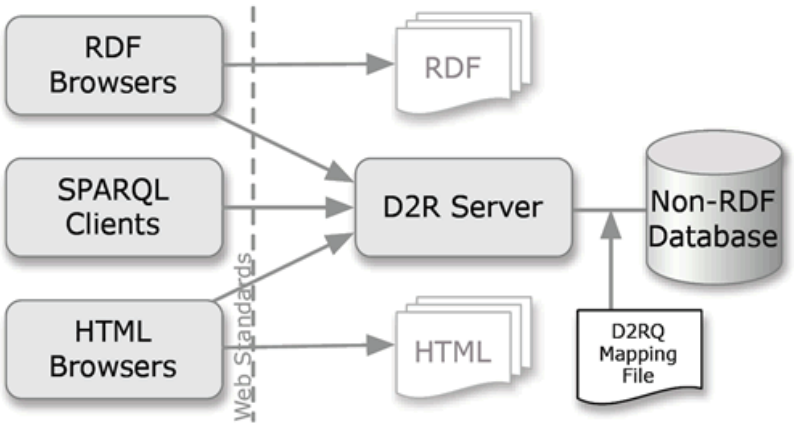

Figure 18 Serveur D2R pour publication de données liées

Il existe plusieurs serveurs D2R en ligne, par exemple Berlin DBLP Bibliographie serveur , Hannover DBLP Bibliographie serveur.

- **Utilisation d'autres types d'informations**

Si les informations sont représentés dans des formats tels que CSV, Microsoft Excel, ou BibTEX et qu'on veut les publier au format données liées sur le Web, il est souhaitable de procéder comme suite :

- Convertir les données en RDF en utilisant des outils comme RDFizing.

- Après la conversion, stocker les données dans un référentiel RDF. Une liste des référentiels RDF est maintenue dans le wiki ESW. Idéalement, le référentiel RDF choisi devrait venir avec une interface de données liées qui prend soin de faire votre Web de données accessible. Comme de nombreux référentiels RDF n'ont pas encore mis en œuvre les interfaces de données liées, on peut également choisir un référentiel qui fournit une extrémité SPARQL et mettre Pubby comme une interface de données liées en face de votre terminal SPARQL.

3.6.5. Utilisation des données liées aux réseaux sociaux

Les données liées peuvent également être utilisés dans les réseaux sociaux et l'utilisation est encore en croissance. Même les grands noms de réseaux sociaux (Facebook, Google Plus) travaillent actuellement à introduire la sémantique aux informations, afin de faciliter le développement et permettre plus de nouvelles applications.

1) Friend Of A Friend (FOAF) : L'ami d'un ami, un projet qui a l'intention de construire un réseau lisible à la fois par la machine et les gens, en les reliant à tout ce qui leur sont liées. FOAF est un *RDF vocubulary* pour décrire les gens, leurs relations et leurs propriétés. Un profil FOAF décrit une personne, quelques faits de cette personne (par exemple, les intérêts, les projets en cours), et des connexions à d'autres personnes (personnes « connus » par cette personne) d'une manière lisible par la machine et défini sémantiquement. Il permet aux gens de créer des pages Web lisibles par la machine pour les personnes, les groupes, les organisations et autres concepts connexes.

2) Open Graph de Facebook : Facebook a lancé récemment un protocole pour accéder à son réseau social. C'est ce qu'on appelle Open Graph et il a maintenant la compatibilité avec RDF qui est une étape intéressante pour faire Facebook partie du web sémantique. Outre Facebook, Google Plus a également mis en place le support du format Open Graph.

3) Communautés en ligne reliées sémantiquement (SIOC) : Les Communautés en ligne reliées sémantiquement est un projet qui fournit une ontologie du Web sémantique pour faire une représentation riche des données en RDF. Un de ses principaux objectifs est de fournir l'interopérabilité entre les différents réseaux sociaux. Il prévoit la création d'un format d'échange qui permet à n'importe quel réseau social d'interagir avec les autres. En outre, il fournit des informations sémantiques sur les réseaux sociaux qui pourraient être très utiles s'elles sont bien appliquées.

3.7. Conclusion

Les Données liées est un concept très puissant qui n'a pas connu jusqu'à présent beaucoup d'évolution. Cette technologie est très importante et utile pour découvrir des nouvelles informations, d'y accéder, de les intégrer et de les utiliser dans des applications grâce à l'utilisation des liens RDF externes qui permet de rendre le web un espace de données interconnectés.

L'utilisation de Linked Data dans les réseaux sociaux ouvre des opportunités aux communautés web pour l'amélioration des échanges d'informations et bénéficier des avantages de cette technologie.

Dans le prochain chapitre nous aborderons les travaux réalisés dans le domaine de recommandation.

Chapitre 4
Systèmes de recommandation

4.1. Introduction

Comme l'indique Chris Anderson dans son ouvrage "the long tail", il semble que "*nous quittons progressivement l'âge de l'information pour rentrer dans l'âge de la recommandation*". Les systèmes de recommandation sont des composants logiciels dont le but est de fournir à des utilisateurs des informations qui correspondent à ces centres d'intérêts et cela on analysant leurs interactions avec leur espace informationnel.

Il s'agit par exemple de déterminer les préférences d'un utilisateur pour lui suggérer des produits personnalisés comme le fait le site de e-commerce Amazon.

Les systèmes de recommandation ont comme finalité de nous aider à traiter des informations dont le volume et la complexité sont en extension continue. Ils se doivent de nous sélectionner les informations les plus intéressantes en fonction des centres d'intérêts des utilisateurs.

Selon leur mode de fonctionnement, les systèmes de recommandation peuvent être *personnalisés* ou *non-personnalisés* [Mohamed Ryadh Dahimene 2014].

Les systèmes de recommandation *non-personnalisés* ne prennent pas en considération les préférences des utilisateurs pour leurs fournir des suggestions.

Les systèmes de recommandation *personnalisés* se basent sur les centres d'intérêts des utilisateurs et leurs interactions avec leur sphère informationnelle afin de fournir des suggestions pertinentes.

4.2. Les systèmes de recommandation personnalisés

Il existe trois grandes approches de systèmes de recommandation *personnalisés* :

- Les systèmes basés sur le contenu
- Les systèmes basés sur le filtrage collaboratif
- Les systèmes hybrides.

4.2.1. Les systèmes basés sur le contenu

Les systèmes de recommandation basés sur le contenu fonctionnent en analysant les caractéristiques des objets à recommander (produits, etc.) puis en les regroupant. Par la suite, le système va suggérer aux utilisateurs ayant acheté/consommé un produit quelconque par le passé, les objets/produits estimés similaires [Ricci et al., 2011].

L'architecture générale d'un système de recommandation basé sur le contenu s'articule autour de 3 modules principaux :

- **L'analyseur de contenu** : ce module d'analyse de contenu a pour objectif de construire une description structurée des objets à recommandés où une étape de pré-traitement est nécessaire afin d'en extraire les caractéristiques. Cette description va servir d'élément d'entrée aux autres modules.

- **Le module d'apprentissage de profils** : ce module construit une description des préférences des utilisateurs on analysant les interactions passées de l'utilisateur sur les objets du système.

- **Le module de filtrage –** en se basant sur les centres d'intérêts des utilisateurs issus de l'étape précédente et sur des descriptions des objets à recommander, ce module construit des listes de suggestions à présenter aux utilisateurs.

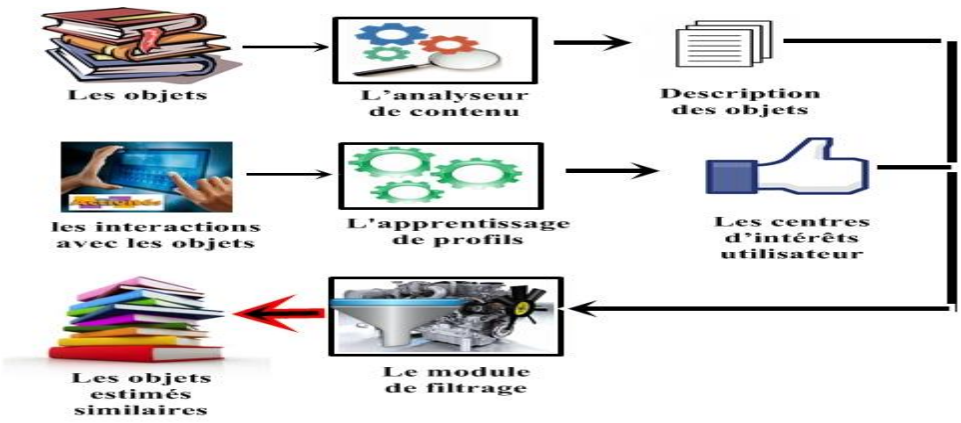

Figure 19 Systèmes de recommandation basés sur le contenu

4.2.2. Le filtrage collaboratif

Le filtrage collaboratif regroupe l'ensemble des méthodes qui visent à construire des systèmes de recommandation utilisant les opinions et évaluations d'un groupe d'utilisateurs pour assister un individu dans son choix.

Un des exemples les plus connus d'un tel système a été développé par le site de commerce en ligne *Amazon.com* et son algorithme de *Item-to-item Collaborative Filtering* qui se traduit sur le site par la fonctionnalité "Les gens qui ont acheté le produit *x* ont aussi acheté le produit *y*" [Linden et al.,2003].

Le filtrage collaboratif a pour avantage qu'elle ne nécessite pas une description précise des objets à suggérer. Ce qui permet de recommander des objets complexes sans avoir à les analyser du fait que les recommandations étant basées sur l'ensemble des interactions des utilisateurs avec les objets du système.

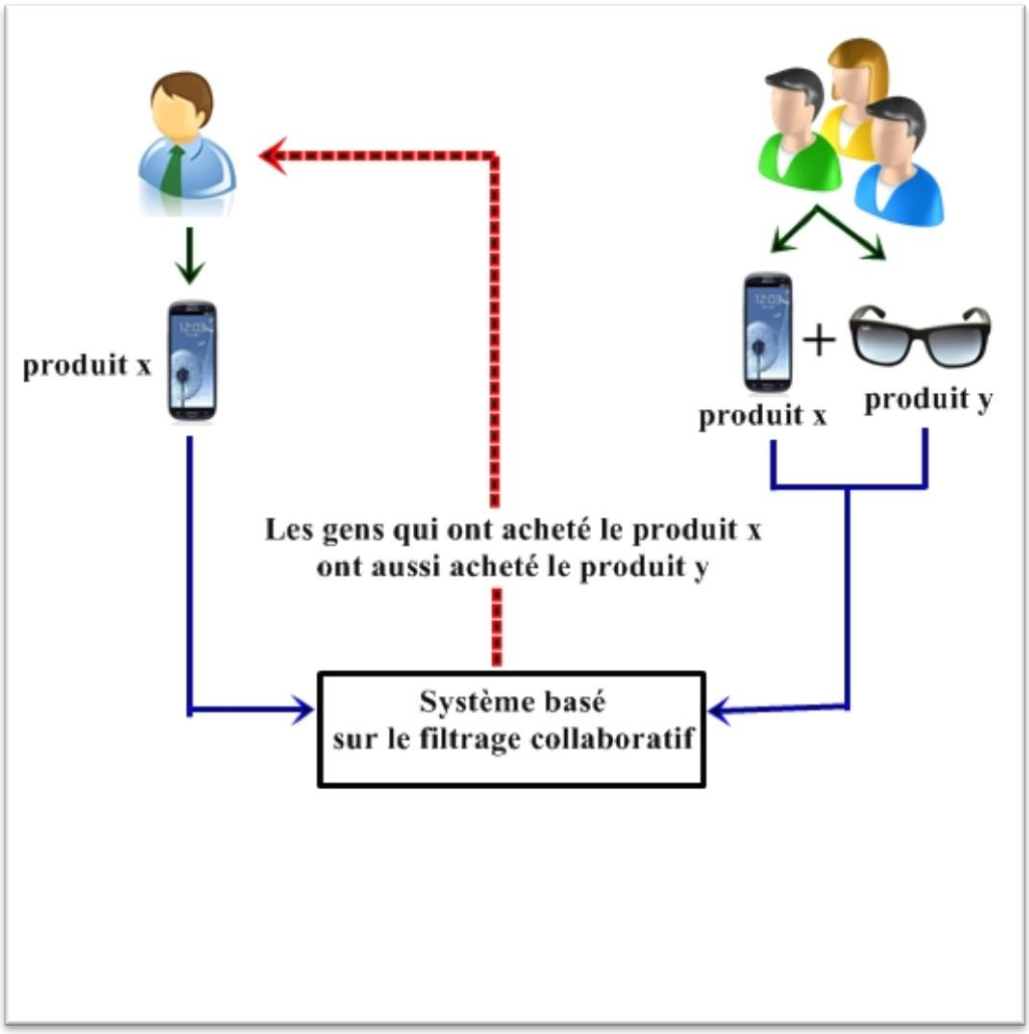

Figure 20 Les systèmes de recommandation basés sur le filtrage collaboratif

4.2.3. Les systèmes hybrides

Les systèmes hybrides permettent de résoudre les problèmes posés par l'utilisation de l'un des deux approches citées ci-dessus. Par exemple la première approche nécessite un riche historique d'interaction avec les objets du système et une description détaillé de ces derniers ce qui n'est pas évident dans certain cas (utilisateur fraichement inscrit). Par contre la deuxième, ont besoin de l'existence d'une large base d'interactions sur l'ensemble du catalogue d'objets du système afin de pouvoir calculer des rapprochements entre les utilisateurs.

La plupart des systèmes de recommandation existants privilégient le modèle hybride au vu des avantages qu'ils présentent. Parmi les systèmes hybrides les plus connus, le système de recommandation mis en place par le géant Américain de la vidéo à la demande sur Internet *Netflix.* [Amatriain, 2013].

4.3. Conclusion

Dans ce chapitre nous avons abordé les différents systèmes de recommandation dans le web. Selon leur mode de fonctionnement, les systèmes de recommandation peuvent être personnalisés ou non-personnalisés.

Nous nous intéressons plus particulièrement aux systèmes de recommandation personnalisées où nous avons détaillé les trois différentes approches basés sur le contenu, le filtrage collaboratif et le système hybride.

Le prochain chapitre « conception et réalisation » représente l'architecture de notre approche, les technologies et les outils utilisés pour l'implémentation et la mise en œuvre de notre application.

Chapitre 5
Conception et Réalisation du projet RecLiv

5.1.Introduction :

L'essor du Web 2.0 est essentiellement dû aux réseaux sociaux numériques tel que Twitter, Facebook, MySpace, LinkedIn, Viadeo, etc. qui ont proposé aux internautes un moyen d'avoir de multiples possibilités d'interactions sociales via Internet.

Ces environnements nous intéressent dans le cadre de notre projet pour la raison majeure : Ils fournissent aux internautes de nombreuses fonctionnalités (tags, mur, photos, liens, groupes, pages, événements, commentaires, applications de parties tierces, etc.) leur permettant de générer un maximum de traces d'activités et d'interactions. On dispose alors d'importantes quantités de données potentiellement utiles pour la construction des profils utilisateurs.

Ces environnements fournissent en général des API à des développeurs tiers leur permettant de proposer de nouvelles fonctionnalités à leurs utilisateurs en exploitant les masses de données produites par ces derniers. Une application tierce *Anniversaire* s'appuiera par exemple sur les dates de naissance des utilisateurs pour proposer un calendrier dans lequel les dates d'anniversaire sont indiquées. La question qui se pose dans notre cas particulier consiste à analyser l'accessibilité à des données de l'utilisateur et de son réseau égocentrique (via les API).

Nous nous intéressons particulièrement au cas de Twitter qui est un réseau social numérique largement utilisé de nos jours d'une part, et qui dispose également d'une API qui est plus riche en terme de fonctionnalités et plus utilisée par les développeurs d'autre part.

Ainsi pour présenter notre première évaluation sur Twitter, nous allons dans un premier temps présenter l'accessibilité aux données utilisateurs qui nous sont utiles via l'utilisation de l'API Twitter. Nous allons ensuite présenter la méthodologie de construction des deux dimensions du profil des utilisateurs à partir des données accessibles. Ensuite le profil utilisateur obtenu sera enrichi avec un jeu de donnée plus important publié en qualité Linked Data. Et enfin en dernière étape nous proposons une recommandation des livres en se basant sur les centres d'intérêts de l'utilisateur.

5.2. Méthodologie de construction et enrichissement du profil utilisateurs

L'architecture de notre approche repose sur cinq phases importantes (figure 21)

1- **Extraction** de données (Accès aux données utilisateurs via l'Api Twitter).
2- **Prétraitement** et **traitement** de données textuelles (pour construire les centres d'intérêts et la structure du réseau égocentrique de chaque utilisateur).
3- **Utilisation** de Linked Data dans le réseau social (utilisation **DBLP**).
4- **Enrichissement** du profil et visualisation de résultat.
5- **Recommandation** des livres publiés en qualité Linked Data.

Chacune de ces étapes est décrite dans les sections qui suivent :

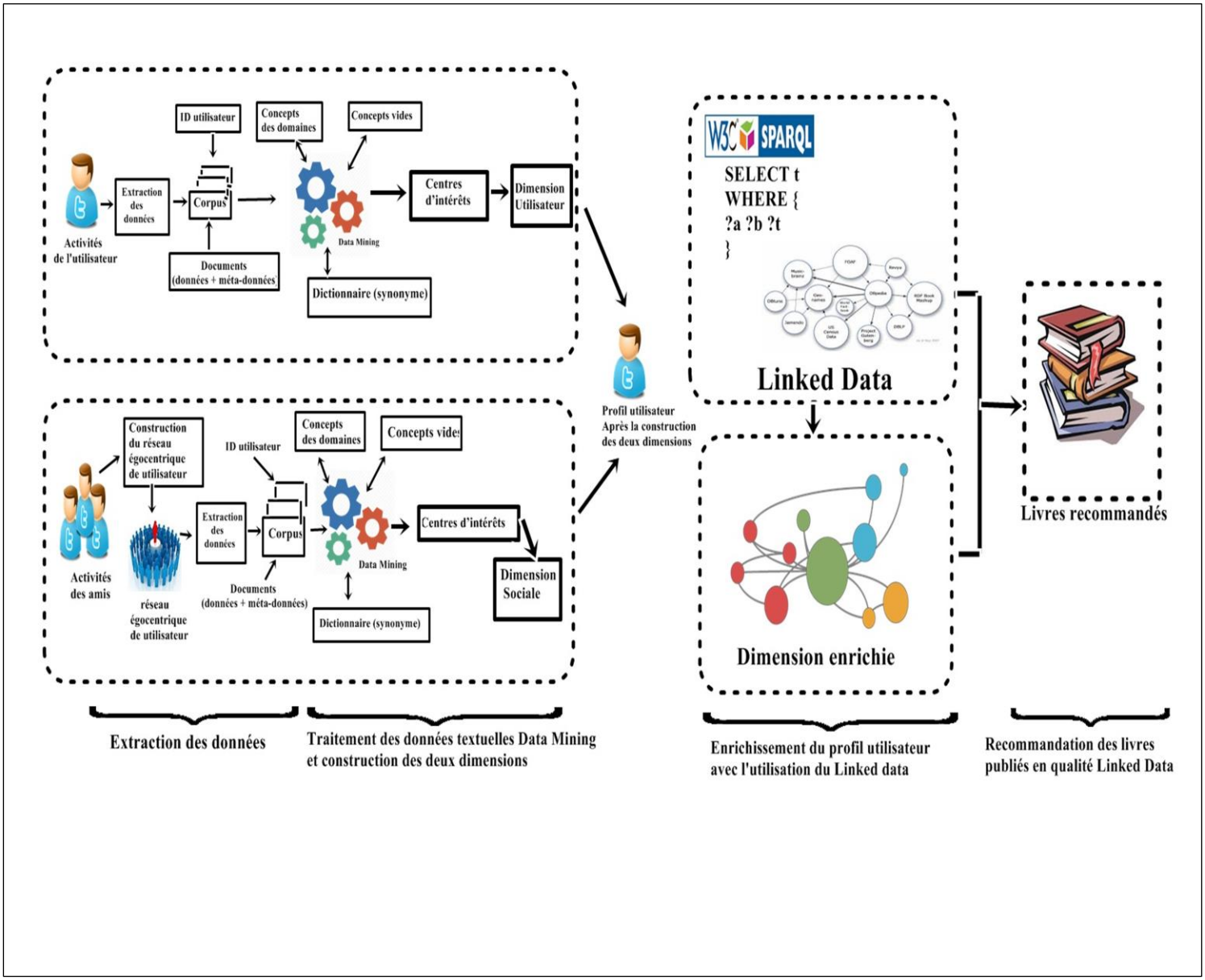

Figure 21 Architecture générale de l'application

5.2.1. Extraction des données

Accès aux données utilisateurs via l'API Twitter :

Généralités sur le développement d'applications Twitter

Twitter est parmi les sites des réseaux sociaux numériques qui ont proposés une API pour le développement de nouvelles fonctionnalités par des tiers. D'un point de vue technique, une application Web développée via l'API Twitter par un tiers est hébergée sur un serveur d'application Web (PHP par exemple) comme toute application Web traditionnelle, même si les interfaces de ces applications sont généralement affichées sur Twitter via le compte de l'utilisateur. En réalité, l'application hébergée sur un serveur d'application tiers interagit avec Twitter via l'API de ce dernier. Pour utiliser une application Twitter développée par un tiers, chaque utilisateur Twitter doit explicitement valider l'installation de cette application sur son profil.

Par rapport aux applications Web traditionnelles, les applications Web développées par des tiers sur Twitter (ou sur les réseaux sociaux numériques en général) ont deux valeurs ajoutées principales. Premièrement, elles peuvent exploiter la structure du graphe social des utilisateurs pour diffuser largement et rapidement des informations. Deuxièmement, elles peuvent accéder à certaines données et à des traces d'activités publiées par les utilisateurs pour personnaliser par exemple les contenus qu'elles leurs proposent.

Pour un utilisateur (*égo*) donne dans Twitter, il s'agit pour nous de rechercher les catégories de données :

- **Les données relatives à l'utilisateur :** ce sont les données r**enseignées explicitement** par l'utilisateur (sexe, date de naissance, cursus académiques, employeurs, etc.), et les données issues des activités de l'utilisateur (tags, commentaires, statuts, liens publies, groupes rejoints par l'utilisateur, événements dont l'utilisateur est un participant, les pages dont l'utilisateur est un fan, etc.). Ces données seront utiles pour construire « *la dimension utilisateur* » du profil utilisateur.

- **Les données de structure du réseau égocentrique de l'utilisateur :** il s'agit des **contacts directs de l'utilisateur** ainsi que les **relations entre ces contacts**. Ces données seront utiles pour construire le graphe représentant le réseau égocentrique de l'utilisateur.

Depuis son apparition l'API Twitter a connue plusieurs changements sur les données accessibles par les applications tierces et sur la manière d'accéder à ces données. Dans un premier temps, lors de la validation de l'installation d'une application tierce sur son profil, l'utilisateur n'avait pas le moyen de paramétrer les données accessibles par cette application à partir de son profil.

Aujourd'hui, cette manière a évoluée et plusieurs « permissions » ont été mises en place pour permettre à l'utilisateur un contrôle de l'accès aux données de son profil par des applications tierces. Certaines données restent tout de même accessibles par défaut dès que l'utilisateur installe l'application sur son profil. Par contre d'autres nécessitent l'accord explicite de l'utilisateur.

Pour réaliser une première évaluation du modèle dans le cas de Twitter, nous avons donc développé une application tierce nommée « *RecLiv* » via l'API Twitter dont le but est de permettre aux utilisateurs volontaires de nous donner les droits d'accès nécessaires à leur profil afin de construire la dimension sociale.

Nous présentons tout d'abord la méthodologie de construction des éléments du profil par l'analyse des données textuelles publiées par les utilisateurs.

5.2.2. Prétraitement des données textuelles

A. Collecte des données :

Les données ou métadonnées extraites sur les activités des utilisateurs sont essentiellement textuelles. Elles sont prétraitées et analysées (fouille de textes). Les données extraites sont représentées sous forme d'un document textuel comprenant les informations sur chaque activité. Les informations d'utilisateurs sont cryptées avec la méthode TF-IDF. Les entités textuelles sont extraites à partir d'un descripteur de document indiquant les séparateurs de texte (blancs, signes de ponctuations, retour à la ligne, etc.). Pour chaque corpus, nous exécutons un prétraitement au nettoyage et filtrons des données.

B. Nettoyage du Corpus :

Dans cette étape, toutes les données obtenues exigent un nettoyage automatique et manuel. On montre l'opération exécutée dans la figure 22.

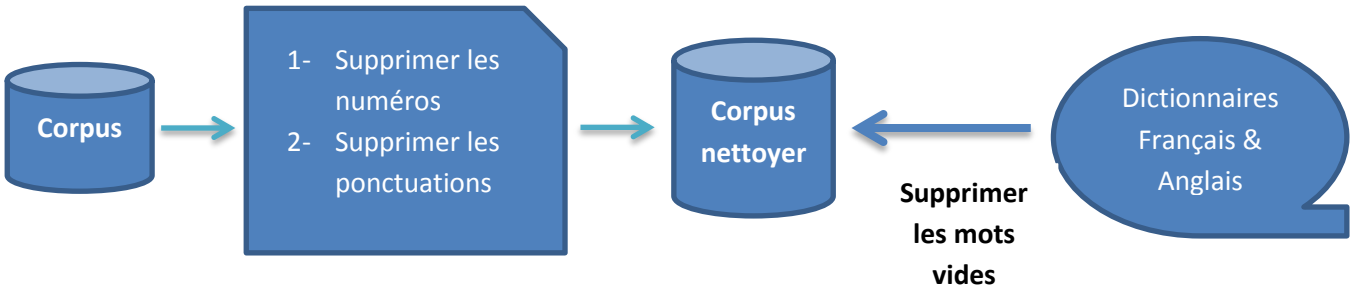

Figure 22 Nettoyage du corpus

Puisque les utilisateurs dans des réseaux sociaux peuvent choisir n'importe quel mot-clé pour catégoriser leur contenu, ils appliquent leurs propres règles d'orthographe (des noms par exemple singuliers ou pluriels, des verbes conjugués). Par conséquent, les termes sont pollués et doivent être nettoyés. Donc, nous devrons nettoyer tous les corpus avant de les analyser.

À cette fin, nous procédons comme décrit dans l'algorithme 1 cité ci-dessous. On commence par **l'élimination des numéros**, et **des ponctuations.** Puis on compare les termes avec les dictionnaires anglais et français afin d'éliminer les mots vides.

Les corpus à nettoyer sont :

- **Corpus utilisateur :** contenant toutes les données personnelles d'utilisateur (les mises à jour de statuts, les messages, les groupes suivis, les personnes suivis, …) sans les données de leurs amis.
- **Corpus amis utilisateur :** contenant toutes les données concernant les amis de l'utilisateur (les mises à jour de statuts pour les amis, les messages des amis, les groupes suivis par les amis, les personnes suivis par les amis, …).

Algorithme 1 : nettoyage du corpus
Entrée : corpus (datasetfile)
Sortie : corpus nettoyé
Tan que non fin (datasetfile) **faire**
Début
Lire la ligne du datasetfile ;
Si le terme est mot vide ou numéro ou ponctuations **alors** Lire la ligne suivante
Sinon voir si le terme existe dans le dictionnaire anglais ou Français
Si le terme n'existe pas dans le dictionnaire **alors** Lire la ligne suivante
Sinon appliquer la stemmatisation au terme
Ajouter le terme au corpus nettoyé
Fin si
Fin Si
Fin

Pour chaque corpus nettoyé, nous calculons le poids de chaque terme en utilisant la technique de TF-IDF.

$$tfidf(t, D) = tf(t, d) \times idf(t, D)$$

TF est le nombre d'occurrence du terme t dans le document d. IDF mesure l'importance du terme t dans le corpus des documents.

$$idf(t, D) = log \frac{|D|}{|d \in D: t \in u|}$$

Où |D | représente le nombre total de documents dans le corpus et $|d \in D: t \in d|$ représente le nombre de documents où le terme t apparaît.
Dans notre cas, le nombre de documents |D | représente le nombre de personne connectée directement à l'utilisateur.

5.2.3. Méthodologie de construction des deux dimensions sociales et utilisateur

A-Construction de la « dimension utilisateur » du profil d'un utilisateur Twitter :

Pour construire **la dimension utilisateur** du profil de l'utilisateur, nous **utilisons** ses **activités** dans **Twitter**. La méthodologie de construction présentée sur la figure 23. Elle est réalisée en trois étapes :

Etape 1 : concerne les connexions de l'utilisateur aux activités **où** plusieurs **éléments publiés** de l'utilisateur dans Twitter **peuvent** être **exploités** pour **construire** ses **centres d'intérêts**. La **description textuelle** est **analysée** pour **extraire** les mots dans le texte en utilisant des délimiteurs de mots (virgules, espaces, point-virgule, etc.).

Etape 2 : repose sur l'utilisation des techniques du Text Mining. Le nombre d'occurrences de chaque mot est calculé en utilisant la mesure TF-IDF. Les mots obtenus suite à l'application de filtres et de dictionnaires seront considérées comme les centres d'intérêts de l'utilisateur.

Etape 3 : construction de la dimension du profil utilisateur constitué des informations personnelles entrées explicitement par l'utilisateur et les centres d'intérêts obtenus dans l'étape précédente.

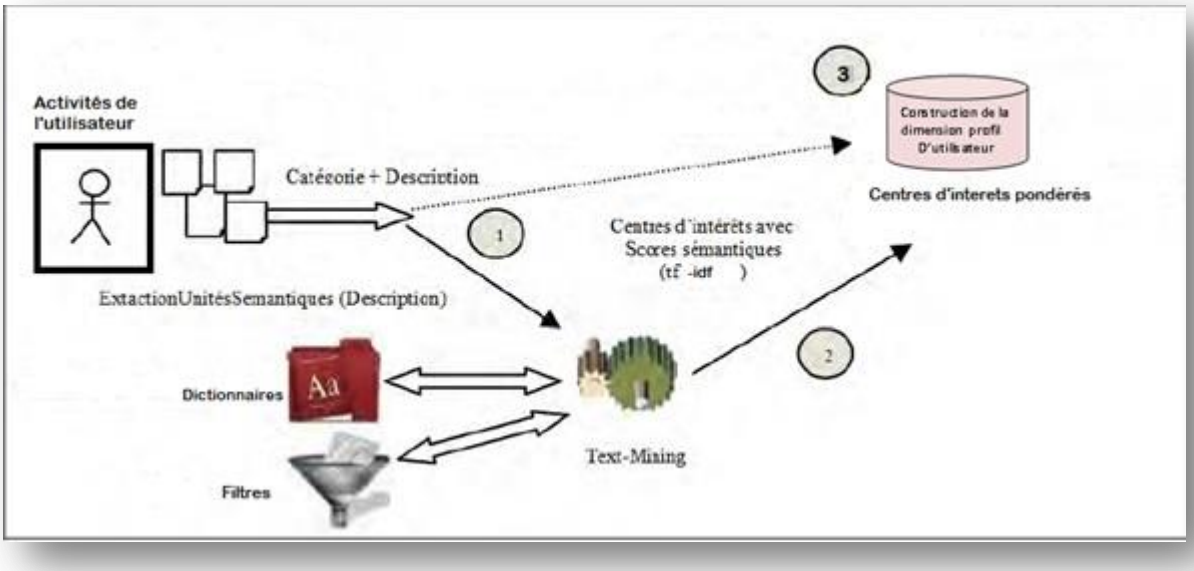

Figure 23 Méthodologie de construction des centres d'intérêts de la dimension utilisateur

B- Construction la dimension sociale du profil d'un utilisateur Twitter

Cette méthodologie se décompose en quatre étapes dont certains sont similaires à celles de la construction de la dimension utilisateur du profil utilisateur.

Etape 1 :

Construction du réseau égocentrique d'un utilisateur :

La notion de réseau égocentrique est principalement utilisée en sociologie. Il s'agit d'un graphe composé des relations entre les individus situés à distance « 1 » (directement reliés) de l'utilisateur (appelé *égo*). Cette notion peut être généralisée pour prendre en compte les utilisateurs situées à distance k de l'*égo* dans le réseau social. Si on considère un réseau social modélisé sous forme d'un graphe *G = (V, E),* V étant l'ensemble des individus, et E l'ensemble des liens entre ces individus. Le réseau égocentrique d'un individu v Є V peut être défini comme :

$Rego(v) = G'(V', E')\ G / \forall u \in V', d(u, v)=k$ et $\exists u' \in V' / (u, u') \in E'$

Pour k=1, le graphe G' correspond au réseau égocentrique tel que défini en sociologie. C'est celui que nous prendrons en compte dans notre proposition. Dans le graphe que nous étudierons, les propriétés d'un nœud sont les attributs de son profil, et nous nous intéresserons particulièrement aux attributs qui représentent les centres d'intérêts de l'utilisateur.

Notre motivation pour l'usage du réseau égocentrique d'un utilisateur est liée au fait que nous considérons que dans la vie réelle, les centres d'intérêts d'un utilisateur sont directement liés aux différentes communautés dont il fait partie. Par exemple, il est normal de considérer qu'un utilisateur qui est inscrit dans un club de tennis, possède le tennis comme centre d'intérêt. Pour retrouver ce centre d'intérêt, nous considérons le fait que les contacts (liens directs) de l'utilisateur qui sont abonnés à ce club de tennis se connaissent également, et possèdent un nombre plus important de liens entre eux, par rapport aux autres contacts de l'utilisateur, qui ne sont pas abonnés à ce club.

Etape 2 : c'est l'extraction des mots, Pour la validation du modèle « social » de profil proposé, nous nous sommes intéressés uniquement à certaines activités des utilisateurs.

Etape 3 : se base sur l'utilisation des techniques du Text Mining. La pondération de chaque mot est calculée en utilisant la mesure TF-IDF. Les mots obtenus suite à l'application de filtres et de dictionnaires seront considérées comme les centres d'intérêts pour le réseau egocentrique de l'utilisateur.

Etape 4 : construction de la dimension social du profil utilisateur utilisant le modèle relationnel avec les attributs de profil social (chaque utilisateur avec leurs intérêts).

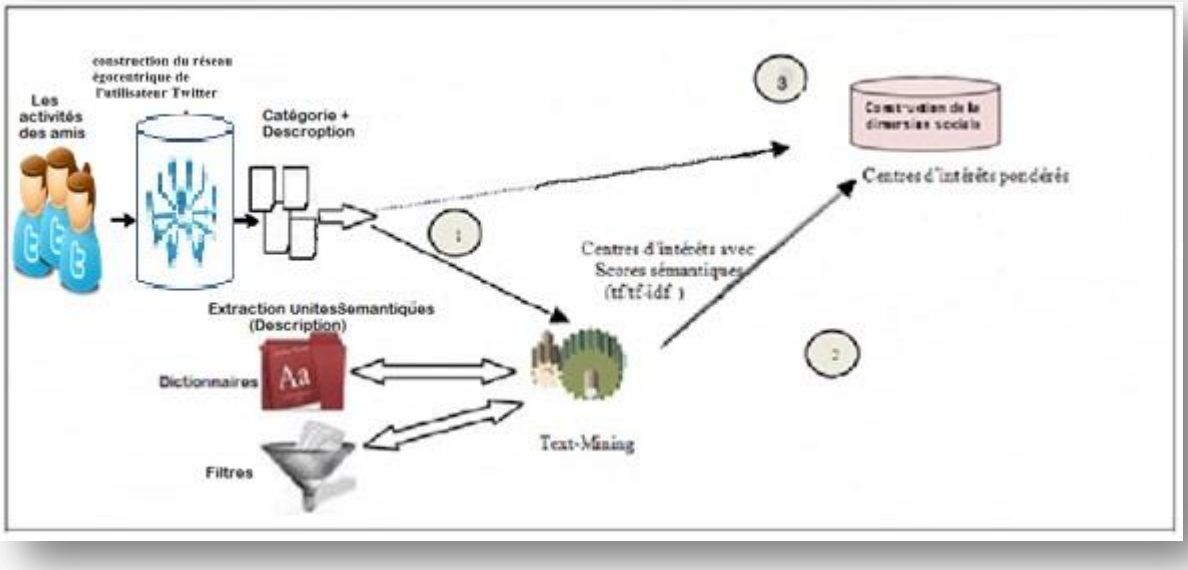

Figure 24 Méthodologie de construction des centres d'intérêts de la dimension sociale

Une fois les centres d'intérêts sont construits pour les deux dimensions sociale et utilisateur, ils seront exploités éventuellement pour interroger Linked data DBLP dans le but d'ultérieures recommandations.

5.2.4. Enrichissement du profil utilisateur avec Linked data

Le but de cette étape est d'enrichir les résultats obtenus dans Twitter par un jeu de données plus important. Nous avons préféré d'enrichir un profil académique. Dans ce qui suit, nous présentons l'accès aux données et la méthodologie d'enrichissement du profil utilisateur avec Linked Data.

A-Accès aux données dans DBLP :

DBLP[2]est l'une des plus importantes librairies digitales qui référencie de nombreux articles scientifiques publiés dans le domaine de l'informatique dans le monde (conférences, journaux, séries, livres). Les données de ce site peuvent être accessibles au téléchargement. Un article scientifique est représenté dans la bibliothèque DBLP par : auteur, titre, année, volume, et des URLs.

[2]http://www.informatik.uni-trier.de/~ley/db/

B-Méthodologie pour enrichir le profil utilisateur avec Linked Data :

Les centres d'intérêts existants dans le profil utilisateur vont être utilisés comme point d'amorce pour interroger les données liées à travers des requêtes SPARQL visant ainsi à trouver des termes qui peuvent représenter des nouveaux centres d'intérêts. Ces derniers permettent l'enrichissement du profil d'utilisateur.

- **Utilisation DBLP locale :**

Nous avons utilisé **SwetoDblp** qui est une grande ontologie peuplée avec un schéma peu profond contenant un grand nombre de données de cas réalistes. SwetoDblp est publiquement disponible en ligne et utilise la ressource communautaire librement disponible.

Nous avons utilisé un extrait de *dataset* de swetoDblp, qui a des données sous forme de fichier RDF. Un exemple de représentation d'un article scientifique de cette *dataset* est présenté dans la figure 25, où l'on peut aisément identifier le fait que l'article en question est publié dans un journal avec les attributs suivant : auteur, titre, les pages correspondantes à l'article dans le journal, l'année de publication, le volume, le numéro, le nom du journal et l'URL associée.

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<!DOCTYPE rdf:RDF [<!ENTITY xsd "http://www.w3.org/2001/XMLSchema#">]>

<rdf:RDF

  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:opus="http://lsdis.cs.uga.edu/projects/semdis/opus#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xml:base="http://lsdis.cs.uga.edu/projects/semdis/opus#" >

<opus:Book_Chapter rdf:about="http://dblp.uni-
trier.de/rec/bibtex/books/acm/kim95/AnnevelinkACFHK95">

<opus:last_modified_date rdf:datatype="&xsd;date">  2002-01-03
</opus:last_modified_date>

<opus:author>   Fishman:Daniel_H   </opus:author>

<rdfs:label>     Object SQL wsmo – A Language for the Design and
Implementation of Object Databases.        </rdfs:label>

<opus:pages>    42-68     </opus:pages>

<opus:year rdf:datatype="&xsd;gYear">   1995   </opus:year>

<opus:book_title>wsmo</opus:book_title>

<dc:relation>http://www.informatik.uni-trier.de/
~ley/db/books/collections/kim95.html#AnnevelinkACFHK95</dc:relation>

</opus:Book_Chapter>

</rdf:RDF>
```

Figure 25 Exemple de fichier RDF de données de DBLP

On peut interroger ce fichier avec des requêtes SPRQL pour extraire des informations spécifiques qui nous seront utiles pour formuler des recommandations des articles.

- **Utilisation DBLP en ligne :**

L'Explorateur RKB, est l'interface humaine principale actuelle à la Base de Connaissance Résistant (RKB). Il rassemble des données de beaucoup des sources bibliographiques et autres dont la structure a été ajoutée pour avoir la possibilité de les interroger par sujet. L'Explorateur RKB peut être lancé à l'adresse URL : http: // rkbexplorer.com/. La page se chargera, avec l'Applet Java qui peut prendre un peu plus de temps. Si la version appropriée de Java n'est pas disponible, l'Explorateur travaillera toujours, même si la vue du réseau graphique ne soit pas disponible. L'interface est maintenant prête à l'emploi. L'utilisateur peut se concentrer sur un élément dans une des sept dimensions :

1-Les gens, 2-Projets, ***3- Mécanismes de Résistance*** ; des métadonnées sur les mécanismes qui ont été étudiés dans le projet ReSIST, ***4-Cours, 5-Publications*** ; Métadonnées à partir d'un large éventail de ressources principalement en Informatique, y compris tous CiteSeer et DBLP, ainsi que de nombreux articles ACM et IEEE, ***6- Organisations*** ; Organisation : académique, gouvernementale, et industrielle, ***7-Domaines de Recherche.***

Pour utiliser cette base de données on procède de la même manière que la méthode précédente d'interrogation de DBLP local, c'est-à-dire obtenir de nouveaux centres d'intérêts à partir de ceux déjà construits dans les deux dimensions utilisateur et social pour enrichir le profil d'utilisateurs.

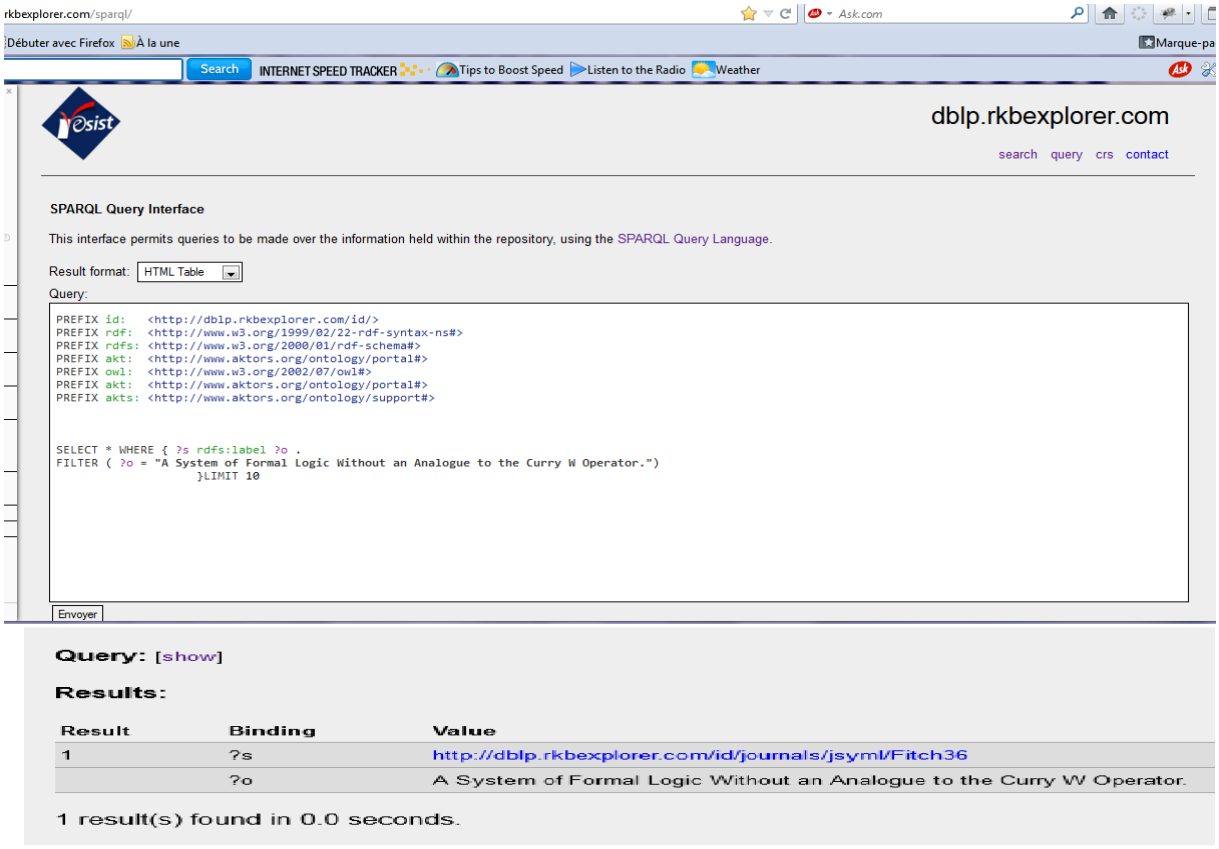

Figure 26 Exemple d'interrogation DBLP en ligne par la Base de Connaissance Résistant (RKB)

Notre implémentation comporte les phases suivantes :

1. Construction du profil d'utilisateur dans un réseau égocentrique avec des ensembles de sujets d'intérêts issus de ses activités et des activités de ses amis.
2. Utilisation des mots clés pertinents à ses sujets d'intérêts pour interroger DBLP en se servant des requêtes SPARQL.
3. Extraction des publications dans DBLP qui traitent ce mot soit dans le titre soit des keywords.
4. Utilisation des nouveaux termes obtenus pour d'enrichir le profil utilisateur.

Le profil de la communauté est obtenu en regroupant tous les centres d'intérêts pondérés des étapes précédentes au sein d'un vecteur unique de centres d'intérêts.

5.2.5. Recommandation des livres publiés en qualité Linked Data

En plus de la partie d'enrichissement d'un profil utilisateur, nous avons aussi développé des requêtes SPARQL pour un système de recommandation qui permet de restituer un ensemble de publications d'articles qui traitent des termes figurants dans les centres d'intérêts des deux dimensions utilisateur et sociale.

Exemple de requête :

```
1  PREFIX  opus:   <http://lsdis.cs.uga.edu/propono#>
2  PREFIX  rdfs:   <http://www.w3.org/2000/01/rdf-schema#>
3
4  SELECT ?label ?author ?volume ?pages ?number
5  WHERE {
6          ?article opus:year ?year .
7          ?article opus:publication_authored_by ?author .
8          ?article rdfs:label ?label .
9          ?article opus:volume ?volume .
10         ?article opus:pages ?pages .
11         ?article opus:number ?number .
12         ?article opus:journal_name ?journal_name .
13
14         FILTER                (?journal_name = "VLDB J.")
15 }
```

Figure 27 Un exemple d'une requête SPARQL sur un data set local DBLP

Le résultat de cette requête est la sélection des informations (le titre , l'auteur, volume, page, nombre) des articles publiés dans un journal nommée 'VLDBJ'.

5.3. Environnement de développement

Nous avons développé notre application sur une machine Intel Core i3 avec une vitesse de 2.27 GHz, dotée d'une capacité mémoire de 4 GB, sous Windows 7. L'application a été réalisée avec le langage de programmation Java sous Netbeans IDE 8.0.2.

Netbeans IDE [3] : est un IDE open source écrit dans un langage de programmation Java. Il fournit des services communs afin de créer des applications de bureau tels que la gestion des fenêtres et de menus, paramètres de stockage…, etc. Le premier IDE a soutenu pleinement les caractéristiques du JDK 5.0. La plate-forme Netbeans IDE est soutenue par Sun Microsystems.

Twitter4J a été créé comme une bibliothèque légère et accessible spécialement développée pour Twitter. Twitter4J est une bibliothèque Java utile et open source.

Pour la création de l'application Twitter on doit procéder comme suite :

- Se connecter à un compte utilisateur en ligne sur Twitter

- Pour une nouvelle application, Aller dans la page « https://apps.twitter.com/» et créer cette application (Figure 28).

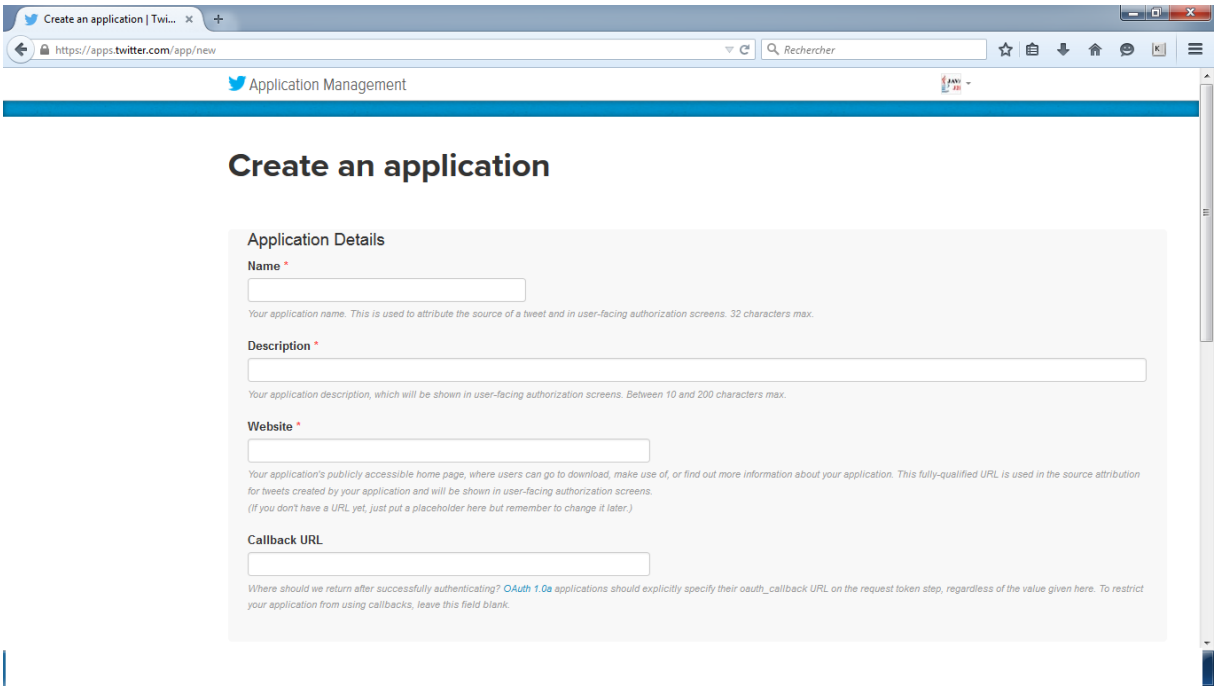

Figure 28 Création d'une application sur Twitter Application

[3] http://www.netbeans.org/

- Utiliser les onglets situé dans l'interface graphique de Twitter developers pour obtenir les codes d'accès avec quelques permissions afin d'utiliser les données du compte de Twitter sur une application (Figure 29).

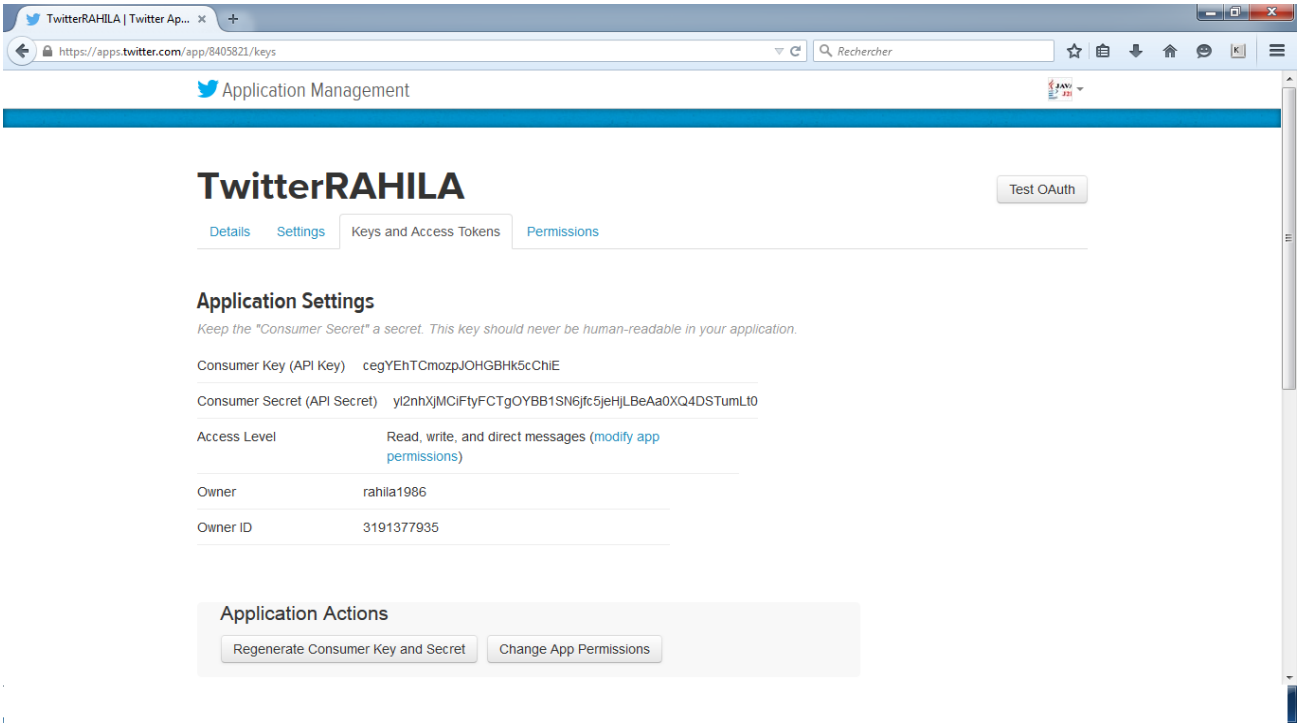

Figure 29 Génération des codes d'accès et modification des permissions

Création de notre Projet RecLiv en Java :

- La première étape de notre travail est la création d'un nouveau projet java sous Netbeans.

- La deuxième étape est d'y intégrer des codes sources indispensables (Figure 30).

- La troisième étape est l'ajout des bibliothèques nécessaires pour interagir avec Twitter et les documents RDF tel que :

- ***Twitter4j*** ayant des méthodes dédiées à Twitter par exemple *getFriendsIDs* pour obtenir les amis d'un compte et *getUserTimeline* pour les statuts publiés,
- ***jena*** qui est une grande bibliothèque prédestinée pour les ontologies (RDF, RDFs, OWL, SPARQL...)

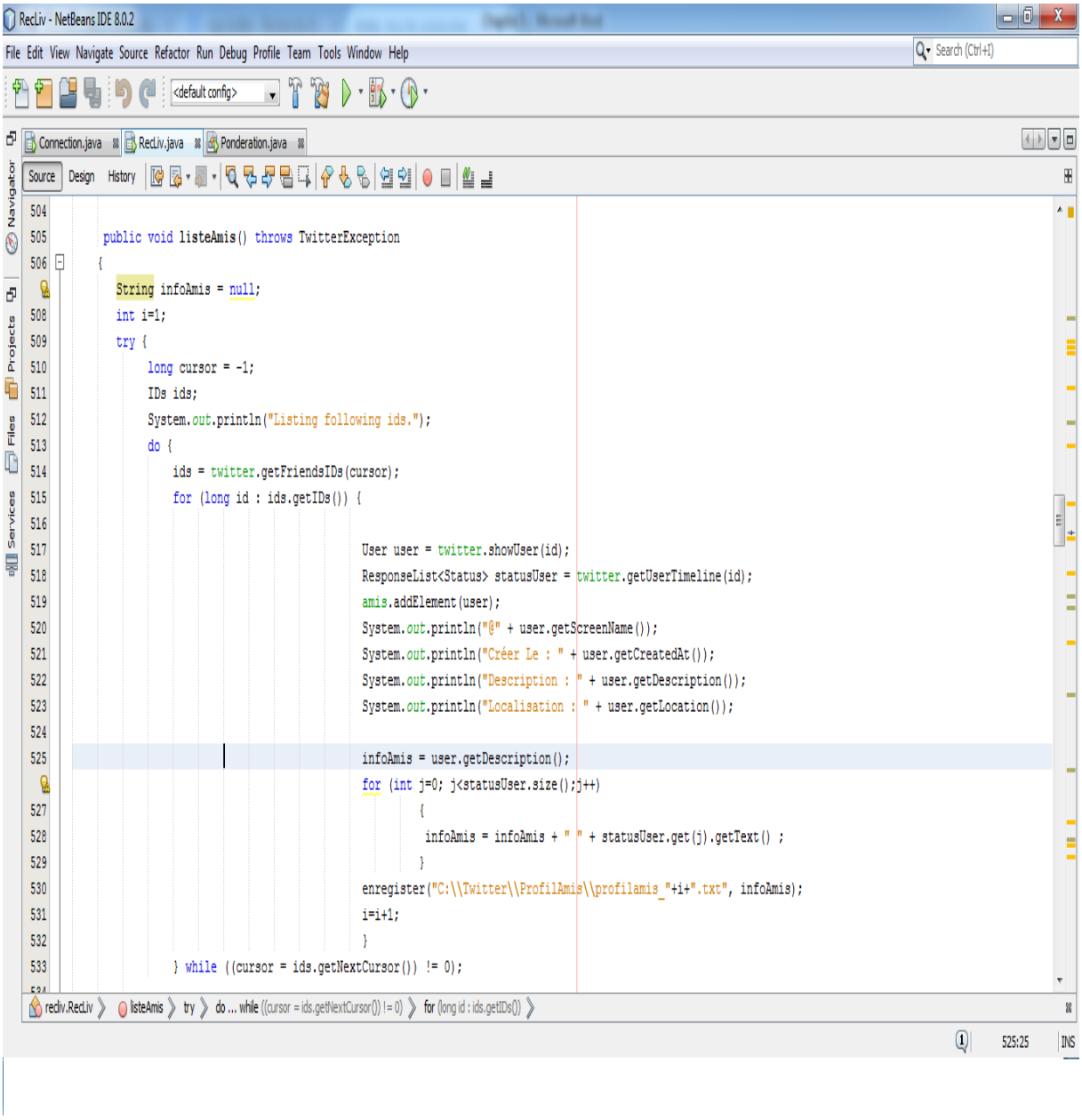

Figure 30 Code source de la méthode « listeAmis » de la class RecLiv

Lancement de notre projet :

Après l'exécution de notre projet RecLiv sous *Netbeans*, une Interface graphique apparait et demande des informations d'authentification pour Twitter (Figure 31).

Figure 31 Interface graphique d'authentification pour se connecter au compte Twitter

Une fois l'authentification est faite une nouvelle fenêtre s'affiche et qui contient tous les méthodes utilisés dans notre approche.

1. **L'onglet Accueil :** représente la page d'accueil de l'application RecLiv. Elle donne les dernières publications du réseau égocentrique de l'utilisateur (Figure 32).

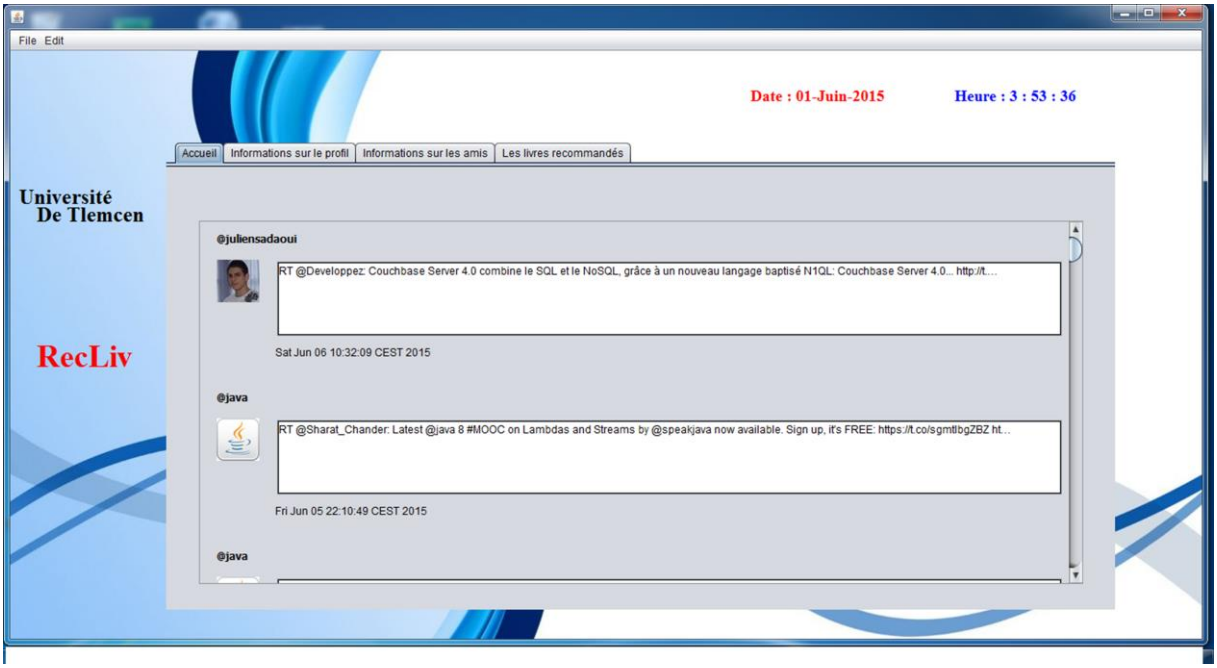

Figure 32 Page d'accueil de l'application RecLiv

2. **L'onglet informations sur le profil :** montre les informations du profil utilisateur et ses intérêts après l'analyse et la construction de la dimension utilisateur (Figure 33).

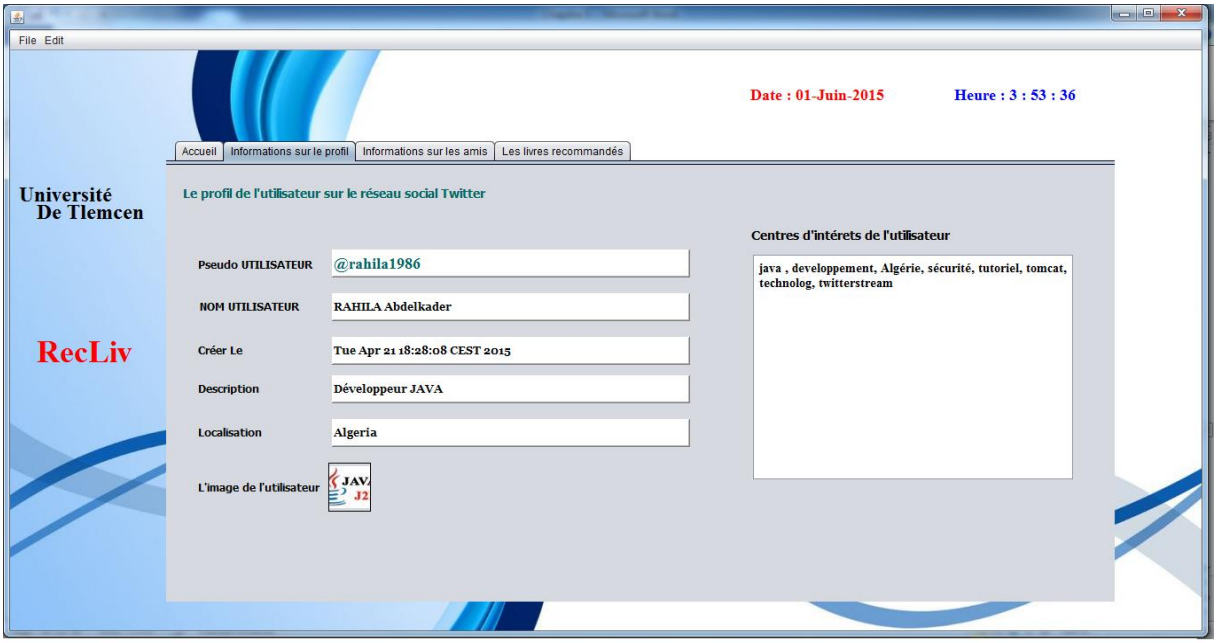

Figure 33 Informations sur le profil utilisateur

3. **L'onglet informations sur les amis :** montre les amis du profil utilisateur sur un réseau égocentrique avec leurs intérêts ce qui implique la construction d'une dimension sociale (Figure 34).

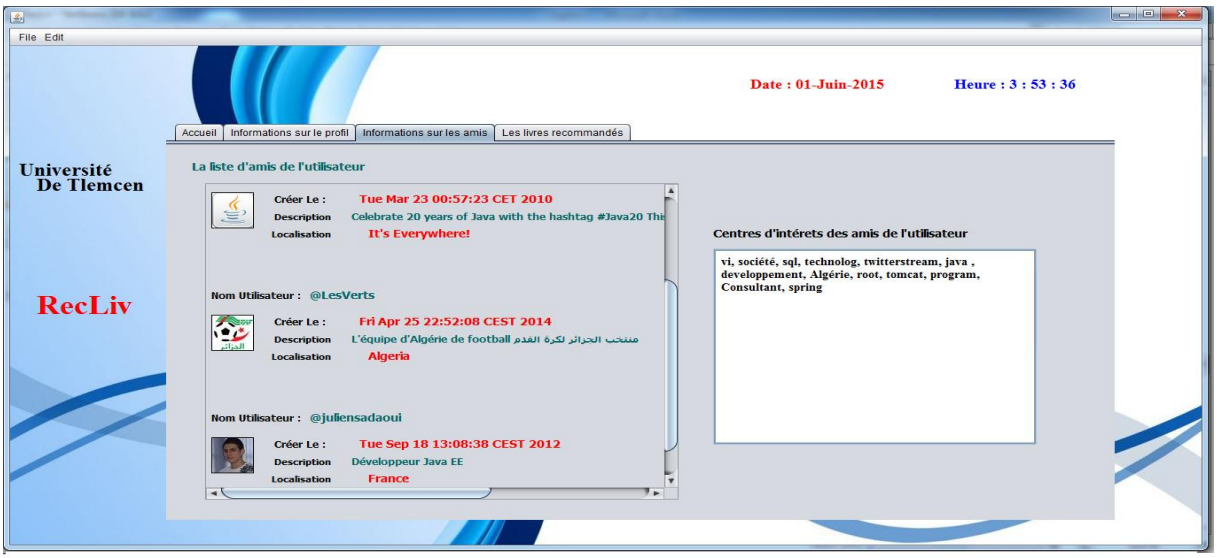

Figure 34 Informations sur les amis de l'utilisateur

Après avoir construit les centres d'intérêts pour les deux dimensions (la dimension utilisateur et la dimension sociale), ces derniers ont été fusionnés pour construire ainsi un vecteur des centres d'intérêts commun. Les éléments de ce vecteur vont être parcouru un par un tout en exécutant des requêtes SPARQL sur le dataset Linked data pour trouver les publications faisant références à ces termes.

Un exemple d'une requête SPARQL est donné dans la figure 35 utilisant les termes stockés dans le vecteur des centres d'intérêts commun.

```
String queryString =

    " PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> " +
    " PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>" +
    " PREFIX opus: <http://lsdis.cs.uga.edu/projects/semdis/opus#>" +
    " PREFIX dc: <http://purl.org/dc/elements/1.1/>" +

    " SELECT ?titre ?label ?auteurs ?nb_p ?year WHERE { " +

    " ?book opus:book_title ?titre ." +
    " ?book rdfs:label ?label . " +
    " ?book opus:author ?auteurs ." +
    " ?book opus:pages ?nb_p ." +
    " ?book opus:year ?year ." +
   " FILTER regex(str(?label), \"" + ponderation.centresInterets[cpt] + "\", \"i\" ) ." +

    " } " ;
```

Figure 35 Requête SPARQL dans notre projet java qui fait un système de recommandation des articles à partir des titres.

4. **L'onglet les livres recommandés :** elle permet d'afficher la liste des livres recommandés pour l'utilisateur.

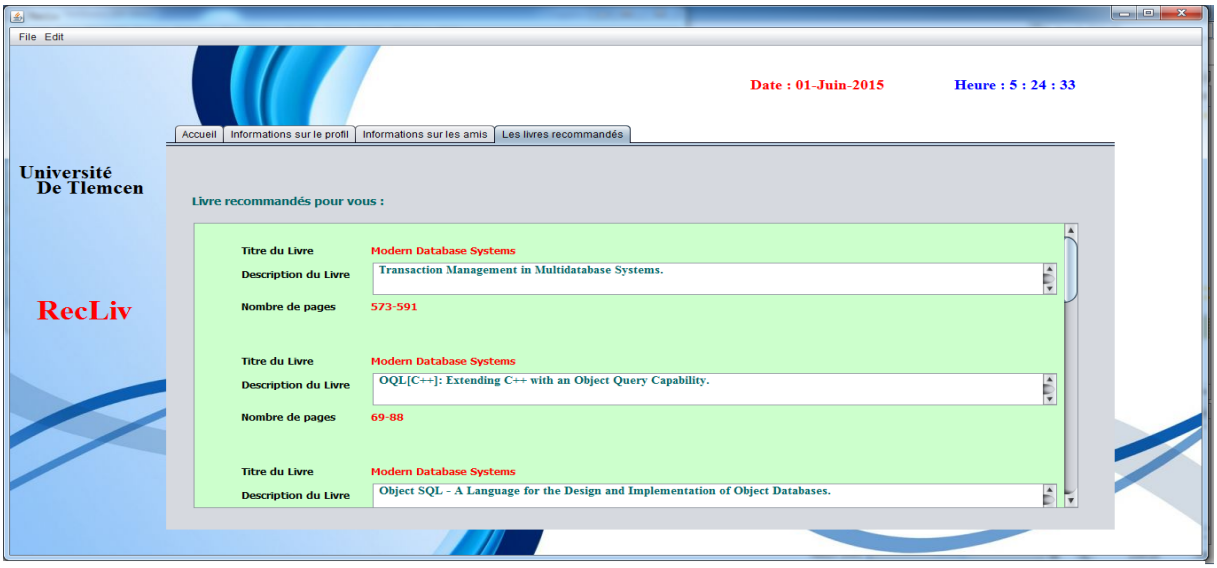

Figure 36 Livres recommandés pour l'utilisateur Twitter

5.4. Conclusion

Dans ce chapitre, nous avons présenté les différentes phases par lesquelles nous sommes passés pour concevoir notre application. En effet, nous avons explicité toutes les étapes pour la construction du profil utilisateur sur un réseau égocentrique via l'API *Twitter*.

En commençant par l'extraction des informations sur les activités de l'utilisateur ainsi que ses amis. Puis à travers l'utilisation des méthodes du Text Mining on a conçu les centres d'intérêts du profil utilisateur.

Ces derniers sont utilisés comme point d'amorce dans les requêtes SPARQL pour interroger les données publiées en qualité Linked Data afin d'enrichir le profil utilisateur et faire éventuellement une recommandation des livres.

Conclusion générale

Le travail présenté dans ce mémoire a pour but d'améliorer l'expérience des utilisateurs des plateformes sociales via l'introduction de deux mécanismes complémentaires qui sont l'enrichissement du profil utilisateur ainsi que la recommandation de livres correspondants à ses centres d'intérêts.

Nous nous sommes intéressés particulièrement au cas de Twitter qui est un réseau social numérique très populaire, disposant d'une part une API qui est plus riche en termes de fonctionnalités et d'autre part plus utilisée par les développeurs.

Dans le contexte de notre projet, nous avons développé une approche qui permet de construire un profil utilisateur constitué d'une dimension utilisateur et d'une dimension sociale. Le résultat obtenu composé d'un ensemble de centres d'intérêts a été enrichi par l'analyse d'un data set de type Linked data.

Le dataset utilisé est celui de DBLP qui a été interrogé pour enrichir la dimension obtenue dans le but de formuler certaines recommandations à l'utilisateur.

Le développement réalisé repose sur l'agencement de plusieurs technologies et langages tels que API Twitter4J, la technique de Data Mining (TF/IDF), SPARQL pour interroger le data set Linked data.

Perspectives :

Notre travail repose principalement sur la construction des profils afin de s'assurer dans un premier temps de la pertinence des profils construits, avant leurs usages par des mécanismes d'adaptation de l'information à l'utilisateur (recherche d'information sociale, systèmes de recommandations sociaux, etc.).

Nous proposons ainsi un profil utilisateur à deux dimensions d'informations (dimension sociale et dimension utilisateur). Pour aller plus loin dans les mécanismes d'adaptation de l'information, on peut s'imaginer plusieurs manières d'utiliser les deux dimensions proposés : par exemple, n'exploiter la dimension sociale que lorsque des informations ne sont pas présentes dans la dimension utilisateur, ou encore d'intégrer systématiquement les informations des deux dimensions dans le mécanisme de filtrage social. Ainsi, pour aller jusqu'au bout des mécanismes, ce serait intéressant de proposer différentes techniques d'usage des deux dimensions du profil en fonction des besoins dans les mécanismes de filtrage social de l'information, et d'évaluer les résultats renvoyés par ces mécanismes en fonction des techniques proposées.

Bibliographie :

- [Anderson, 2006] The Long Tail : Why the Future of Business Is Selling Less of More. Hyperion.

- [Adomavicius & Tuzhilin, 2005] Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng., 17(6):734–749.

- [Amit Sheth Ramesh Jain 2007], Social Networks and the SemanticWeb. Amsterdam: Springer Science+Business Media, 2007.

- [A. James O'Malley 2008] James O'Malley Æ Peter V. Marsden, "The analysis of social networks", Health Serv Outcomes Res Method, pp. 222-269, 2008.

- [Ali Kourtiche & al. 2012] Ali Kourtiche , Amar Bensaber Djamel, Sidi Mohamed Benslimane 'Egocentric social networks analyses to create User profile on facebook', Higher National School of Computer Algiers , University Djilali Liabes Sidi Bel Abbes, Algeria 2012

- [Amatriain, 2013] Amatriain, X. (2013). Big & personal : data and models behind Netflix recommendations. In Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining : Algorithms, Systems, Programming Models and Applications, BigMine 2013, Chicago, IL, USA, August 11, 2013, pages 1–6.

- [Berners-Lee 2000] "Weaving the Web". San Francisco,CA : HarperSanFrancisco

- [Breslin, J., & Decker, S. 2007] The Future of Social - The Need for Semantics. IEEE Internet Computing, 5, 86-90.

- [Chris Bizer. 2007] How to Publish Linked Data on the Web. [Online]. http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/

- [Danah m. boyd 2007] Danah m. boyd. (2007, decembre) wiley online librery. [Online].http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00393.x/full

- Dieudonné TCHUENTE , Nadine BAPTISTE-JESSEL , Marie-Françoise CANUT, " Conception de profils visuels d'utilisateurs à partir de réseaux égocentriques (cas de facebook) ", Institut de Recherche en Informatique de Toulouse

- [Diederich & Iofciu, 2006] Diederich, J. and Iofciu, T. (2006). Finding communities of practice from user profiles based on folksonomies. In Proceedings of the EC-TEL06 Workshops, Crete, Greece , October 1-2, 2006.

- [Engestrom, J 2005] Why some Social Networks Work and Others dont or the Case for Object-centered Sociality. [Online]. http://www.zengestrom.com/blog/2005/04/why-somesocial-network-services-work-and-others-dont-or-the-case-for-object-centered-sociality.html

- [Garrett 2005] Garrett, J. J., (2005) Ajax: A New Approach to Web Applications". Adaptive Path. [Online]. http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications

- [Harry Halpin, 2012] Social semantics: the search for meaning on the web (Vol. 13). Springer Science & Business Media.

- [John G. Breslin & al. 2011] John G. Breslin. Alexandre Passan. Denny Vrande, Social Semantic Web. Berlin Heidelberg: Springer-Verlag, 2011.

- [Jens Lehmann & al. 2011] Jens Lehmann, and Axel-Cyrille Ngonga Ngomo Sören Auer, "Introduction to Linked Data,and Its Lifecycle on the Web," Springer-Verlag, pp. 1-75, 2011.

- [Jérôme Euzenat François Scharffe 2011], "Méthodes et outils pour lier le web des données," INRIA & LIG, 2011.

- [John Domingue & al. 2011] Domingue, J., Fensel, D., & Hendler, J. A. (Eds.). Handbook of semantic web technologies (Vol. 1). Springer Science & Business Media.

- [Linden & al., 2003] Linden, G., Smith, B., and York, J. (2003). Industry report : Amazon.com recommendations : Item-to-item collaborative filtering. IEEE Distributed Systems Online, 4(1).

- (2013) Linked Data - Connect Distributed Data across the Web. [Online]. http://linkeddata.org/faq

- [Luciano Floridi 2009], "Web 2.0 vs. the Semantic Web: A Philosophical Assessment," Research Chair in Philosophy of Information and GPI, pp. 01-04, 2009.

- [L. Getoor & C. P. Diehl 2005], "Link mining: a survey," ACM SIGKDD Explorations Newslette, pp. 3-12, 2005.

- [Mark Fischetti & al. 1999] Mark Fischetti, and Michael L. Dertouzos. Tim Berners-Lee, Weaving the Web : The Original Design and Ultimate Destiny of theWorldWideWeb by its Inventor. San Francisco: Harper , 1999.

- [Maciej Janik & al. 2011] Maciej Janik,Steffen Staab Andreas Harth, "Semantic Web Architecture," Handbook of Semantic Web Technologies, pp. 44-71, 2011.

- [Mohamed Ryadh Dahimene 2014], “Filtrage et Recommandation sur les Réseaux Sociaux” Laboratoire CEDRIC – Équipes ISID/VERTIGO, Paris, Décembre 2014

- [P. Mika, M. Greaves 2008], "SemanticWeb andWeb 2.0," Web Semantics Sci Serv AgentsWorldWideWeb, p. 01, 2008.

- [PETER MIKA 2005], "flink semantic web technology for the extraction and analysis of social network " science direct, mai 2005.

- [Ricci & al., 2011] Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors (2011). Recommender Systems Handbook. Springer.

- [Richard Cyganiak & al. 2009] Richard Cyganiak, and Tobias Gauss Christian Bizer, "The rdf book mashup: From web apis to a web of data," In Proceedings of the Workshop on Scripting for the Semantic Web, 2009.

- [R. Moats. 2007] URN Syntax. [Online]. http://tools.ietf.org/html/rfc2141

- [Ralf Heese & al. 2005] Ralf Heese, Malgorzata Mochol, Radoslaw Oldakowski, Robert Tolksdorf, and Rainer Eckstein Christian Bizer, "the impact of semantic web technologies on job recruitment processes," in Internationale Tagung, 2005.

- [Romain Blin,Julien Subercaze Henry Story 2012], "Turning a Web 2.0 Social Network into a Web 3.0," WWW 2012 – Demos Track, pp. 01-02, 2012.

- [S. Wasserman & K. Faust 1999],"Social Network Analysis: Methods and Applications," Cambridge university presse, vol. 01,1999.

- [sioc-project]. [Online]. http://sioc-project.org/

- [Tim O'Reilly 2005] O'Reilly Network : What Is Web 2.0 : Design Patterns and Business Models for the Next Generation of Software. [Online]. http://www.oreillynet.com/lpt/a/6228

- [Thierry Wellhoff 2012], "Tout ce que vous avez toujours sur les médias sociaux". 8 rue Fourcroy -75017 Paris: Wellcom, 2012.

- Tim O'Reilly is founder and CEO of O'Reilly Media. technology publisher. [Online]. http://oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1

- [Tim Berners-Lee 2005] IswcPodcast, "Tim Berners-Lee Interview at ISWC 2005," in 4th International Semantic Web Conference, Galway Ireland , 2005.

- [Tom Heath & Christian Bizer 2011], Linked Data: Evolving the Web into a Global Data Space. Berlin: Morgan Claypool, 2011.

- [Tim Berners-Lee 2006]. (2006, July) Linked Data. [Online]. http://www.w3.org/DesignIssues/LinkedData.html

- [T. Berners-Lee. 1999] Hypertext Transfer Protocol -- HTTP/1.1. [Online]. Hypertext Transfer Protocol

- [Tom Heath & al. 2009] Tom Heath, and Tim Berners-Lee Christian Bizer. Int. J. Semantic Web Inf. Syst., 5(3):1–22,. [Online]. http://dx.doi.org/10.4018/jswis

- [w3c 2013] w3c. (2013) SEMANTIC WEB. [Online]. http://www.w3.org/standards/semanticweb/data

- [Yinuo Zhang & al. 2013] Yinuo Zhang, Hao Wu, Vikram Sorathia and Viktor K. Prasanna, « Event Recommendation in Social Networks with Linked Data Enablement», University of Southern California, Los Angeles, CA, USA, 2013

AJAX : Asynchronous JavaScript and XML

API: Application Programming Interface

DBLP: Digital Bibliography & Library Project

FOAF: Friend of a Friend

HTML : Hypertext Markup Language

LOD : Linked Open Data

OWL: Ontology Web Langage

RDF: Resource Description Framework

RDF(S): Resource Description Framework Schema

RSS: Really Simple Syndication

RKB: ReSIST Knowledge Base

SNA: Social Network Analysis

SIOC: Semantically-Interlinked Online Communities

SKOS : Simple Knowledge Organization System

SPARQL : SPARQL Protocol and RDF Query Language

TF: Term Frequency

TF-IDF: Term Frequency-Inverse Document Frequency

URI : Uniform Resource Identifiers

URL: Uniform Resource Locator

W3C: World Wide Web Consortium

WWW: World Wide Web

XML : eXtensible Markup Langage

Résumé:

Au cours des dernières années, les services de réseaux sociaux ont gagné en popularité. Ils nous permettent une forte exploration et un partage de nos résultats de manière pratique. Nous nous intéressons au défi de fournir à l'utilisateur une expérience de qualité sur le Web social. L'objectif de ce projet est d'étudier les challenges liés à la forte utilisation du web social et de l'apport du web sémantique et du Linked Data pour mieux appréhender les problématiques des réseaux sociaux. Nous utilisons les données liées pour collecter les informations contextuelles relatives aux utilisateurs et de construire un profil amélioré pour eux. Comme ressource fiable, les données liées sont utilisés pour détecter les connaissances structurées et les différents liens entre elles. Comme étude de cas, nous nous intéressons particulièrement au cas de Twitter qui est un réseau social numérique très populaire, disposant d'une part une API qui est plus riche en termes de fonctionnalités et la plus utilisée par les développeurs d'autre part. Nous utilisons le dataset de type Linked Data pour la recommandation des livres en se basant sur le profil enrichi de l'utilisateur et ces centres d'intérêts.

Mots-clés : Recommandation, Données liées, Réseaux sociaux, fouille de données, Twitter, RDF, SPARQL

Abstract:

In recent years, social networking services have gained phenomenal popularity. They allow us to explore the world and share our findings in a convenient way. We are interested in the challenge of providing the user with a quality experience on the social Web. The objective of this project is to study the challenges related to the high use of the social web and the contribution of the Semantic Web and Linked Data to better understand the problems of social networks. We use the Linked data to collect contextual information about users and build an improved profile for them. As reliable resource Linked data are used to detect structured knowledge and the links between them. As a case study, we are particularly interested in the case of Twitter, which is a very popular digital social network, with offered an API that is richer in features and most used by developers. We use the dataset of type Linked Data for recommending books based on the profile enriched of the user and these interests.

Keywords: Recommendation, Linked data, social networks, Text Mining, Twitter, RDF, SPARQL

الملخص :

في السنوات الأخيرة، اكتسبت خدمات الشبكات الاجتماعية شعبية كبير. بحيث انها تسمح لنا باستكشاف قوي وتقاسم لنتائجنا بطرق عملية .و نحن مهتمون بالتحدي المتمثل في تزويد المستخدم بتجربة ذات جودة على الشبكة الاجتماعية. الهدف من هذا المشروع هو دراسة التحديات المتعلقة بالاستخدام الكثيف للشبكة الاجتماعية ومساهمة الويب الدلالي و "البيانات المرتبطة" للحصول على فهم أفضل لمشاكل الشبكات الاجتماعية. نحن نستخدم البيانات المرتبطة لجمع معلومات سياقية حول المستخدمين وبناء البيانات الشخصية لهم. وتستخدم البيانات المرتبطة كمورد موثوق به، للكشف عن المعرفة المنتظمة والروابط المختلفة فيما بينها. و كدراسة حالة، نحن مهتمون بشكل خاص بتويتر، والذي هو عبارة عن شبكة اجتماعية رقمية شعبية جدا، و لتوفره من ناحية على مكتبة التي هي أكثر ثراء من حيث الميزات والأكثر استخداما من قبل المطورين من ناحية اخرى. نستخدم مجموعة البيانات ذات نوع "البيانات المرتبطة" للتوصية الكتب بناء على ملف تعريف المستخدم واهتماماته.

كلمات مفتاحية : التوصية، البيانات المرتبطة ، الشبكات الاجتماعية، استخراج البيانات، تويتر، اردف، سباركل