CETE Méditerranée

Université Montpellier 2
CNRS / LIRMM

Publication de données d'offre de Transport Collectif sur le web des données

Rapport de stage

Octobre 2011

NOTICE ANALYTIQUE

Organisme commanditaire : CETE Méditerranée, DGITM/AFIMB		
Titre : Publication de données d'offre de Transport Collectif sur le web des données		
Sous-titre :	**Langue** : Français	
Organismes auteurs **Université Montpellier 2, CNRS/LIRMM**	**Rédacteurs ou coordonnateurs** Julien PLU, Patrick GENDRE, François SCHARFFE	**date** octobre 2011

Résumé :

Le web des données autrement appelé « web sémantique » a pour but d'étendre aux données le web 'classique' constitué par le réseau des pages HTML : l'idée est de tirer parti de manière plus 'intelligente' de l'information contenue dans les données et non plus seulement de mettre à disposition des documents reliés entre eux. Comme cela s'est produit pour le web de l'information, on peut espérer un 'effet réseau' de plus en plus important à mesure que des données seront publiées et reliées entre elles.

Le web de données est très lié à la notion d'***open data***, qui commence à se concrétiser dans le domaine des transports publics. En effet, le web des données implique que les données soient mises à disposition (librement) des utilisateurs (tout comme le sont les liens vers le contenu des pages html).

Le parallèle intuitif que l'on peut faire entre navigation dans les données et navigation dans les réseaux multimodaux de transport fait que l'idée d'un web de l'info transport semble naturelle, et une piste à creuser pour améliorer l'interopérabilité de l'information multimodale, objectif de l'Agence Française de l'Information Multimodale et de la Billettique (AFIMB) créée par le ministère du développement durable et des transports en 2010. D'où ce travail exploratoire en vue de comprendre comment les idées du web sémantique pourraient s'appliquer à l'information transport.

L'objectif de ce travail exploratoire est de démontrer comment publier des données concernant le transport collectif (TC) sur le web sémantique, en suivant la démarche du projet Datalift sur la mise en œuvre du web sémantique, financé par l'ANR en 2010, et auquel participent entre autres l'Insee et l'IGN. Ce travail a été réalisé dans le cadre d'un stage par Julien Plu, étudiant en 1ère année Master Informatique de l'université de Montpellier 2 en juin, juillet et août 2011 suite aux premiers contacts entre le CETE et François Scharffe, du LIRMM[1], le coordinateur du projet Datalift, et grâce au soutien de la DGITM et de l'AFIMB.

La faisabilité de publier des données TC sur le web des données est étudiée concrètement à partir de 2 sources de données : l'annuaire des services d'information www.passim.info et les fichiers XML au format Neptune des lignes de transport public des Conseils Généraux de l'Isère et de la Gironde.

Outre une introduction, ce rapport téléchargeable sur le site web du CETE comprend 3 parties centrales : tout d'abord, la présentation de la chaîne complète et des principaux types d'outils permettant de passer de données structurées à leur publication sur le web sémantique, puis leur mise en pratique dans les deux parties suivantes, décrivant comment nous avons procédé pour publier 2 types de données : d'une part l'annuaire des services d'information Passim, d'autre part la description de l'offre théorique d'un réseau de transport public Neptune. Il se termine par une conclusion proposant des pistes pour continuer, et en annexe, par une liste des principales références.

Ce travail exploratoire aura permis :

- de présenter les concepts autour du web des données et l'intérêt potentiel pour la réutilisation de données relatives aux TC ;
- d'identifier deux sources de données potentiellement pertinentes pour faire émerger le web des données de transport public ;
- montrer la faisabilité de mise en oeuvre d'une chaîne de publication complète.

Nous n'avons malheureusement pas pu aller jusqu'à produire une démonstration d'utilisation de ces données pour une application concrète, à la fois par manque de temps, et par manque d'une 'masse critique' de données auxquelles se relier. On peut néanmoins se faire une idée des applications en observant ce que font nos collègues anglais, qui sont les plus avancés en la matière.

Toutefois, compte tenu de la dynamique actuelle en matière d'open data et de réutilisation des données publiques y compris dans le transport, qui sont des conditions nécessaires pour le web sémantique, et de la prochaine ouverture de data.gouv.fr qui permettra a minima de publier le contenu de passim, il nous semble que ce premier travail peut être poursuivi dans plusieurs directions :

- à court terme, nettoyer les erreurs résiduelles dans le fichier RDF de passim ;
- envisager dans le cadre d'une initiative open data d'une collectivité locale AOT qui inclurait un volet 'web

[1] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier

sémantique', une publication de données d'offre TC Neptune ; - par la même occasion, développer des applications significatives qui illustrent les utilisations possibles de ces technologies, ce que nous n'avons pas pu faire lors de ce stage trop court ; - à moyen terme, étendre l'ontologie Neptune à d'autres éléments de données que les lignes et arrêts, et sans doute plutôt au niveau européen en cohérence avec Netex ; - faire le lien avec les projets nationaux de l'AFIMB (référentiels régionaux de points d'arrêts, standardisation des Web services).			
Mots clés : web des données, web sémantique, linked data, transmodel, passim		**Diffusion :** Version électronique	
Nombre de pages : 24 pages		**Confidentialité :** Non	**Bibliographie** : Oui

Mise à jour du 04/10/11

Sommaire

Remerciements.

Nous remercions Michel Girard, de la Direction des Transports du Conseil Général de l'Isère, qui nous a donné accès aux données de TransIsère au format Neptune ainsi que Pascal Romain, du Conseil Général de la Gironde, qui a publié les données de TransGironde au même format, en open data.

I. INTRODUCTION

A. Qu'est-ce que le Web sémantique

Le Web sémantique (appelé aussi « le Web des données ») vise à permettre aux machines d'utiliser la sémantique, c'est-à-dire la signification de l'information, sur le Web. Il étend le réseau des hyperliens entre des pages Web classiques par un réseau de liens entre données structurées, permettant ainsi à des agents automatisés d'accéder plus intelligemment aux différentes sources de données contenues sur le Web et, de cette manière, d'effectuer des tâches (recherche, apprentissage, etc.) plus précises pour les utilisateurs. Le terme a été inventé par Tim Berners-Lee, co-inventeur du Web et directeur du W3C, le consortium international qui supervise l'élaboration des propositions de standards du Web.

Objet de travaux de recherche depuis une bonne dizaine d'années, le web sémantique commence à devenir une réalité avec la publication de plus en plus de données 'open data', par les pouvoirs publics, en vue de favoriser leur réutilisation. Depuis peu, les principaux éditeurs de moteurs de recherches (Microsoft, Yahoo et Google) développent des outils permettant d'ajouter des informations sémantiques aux pages HTML, en vue d'améliorer la pertinence des moteurs de recherche.

B. Contexte

1. Le CETE Méditerranée

Le CETE (Centre d'Etudes Techniques de l'Equipement) Méditerranée est un service technique du ministère du développement durable. Le stage est proposé par le service DCEDI/TIM qui travaille dans le domaine de l'ingénierie du trafic et des systèmes d'information sur les transports. Ses activités sont décrites sur les pages http://www.cete-mediterranee.fr/tt13/www/.

Le CETE contribue aux programmes nationaux relatifs au développement des systèmes d'information sur tous les modes de transport (www.predim.org), et dans ce cadre, en particulier à la mise en oeuvre d'un annuaire des sources d'information transport www.passim.org et d'un logiciel libre www.chouette.mobi permettant la validation, l'échange et l'édition de données décrivant l'offre de réseaux de transport public conformément à un profil d'échange normalisé (XML Neptune).

2. Le web des données pour le transport

Le web des données autrement appelé « web sémantique » a pour but de faire avec les données ce qui existe dans le web de l'information (constitué par le réseau des pages HTML) : l'idée est de tirer parti de manière plus 'intelligente' de l'information contenue dans les données et non plus seulement de mettre à disposition des documents reliés entre eux. Comme cela s'est produit pour le web de l'information, on peut espérer un 'effet réseau' de plus en plus important à mesure que des données seront publiées et reliées entre elles.

Le web de données est très lié à la notion d'***open data***, qui commence à se concrétiser dans le domaine des transports publics. En effet, le web des données implique que les données soient mises à disposition (librement) des utilisateurs (tout comme le sont les liens vers le contenu des pages html).

Le parallèle intuitif que l'on peut faire entre navigation dans les données et navigation dans les réseaux multimodaux de transport fait que l'idée d'un web de l'info transport semble naturelle, et une piste à creuser pour améliorer l'interopérabilité de l'information multimodale, objectif de l'Agence Française de l'Information Multimodale et de la Billettique (AFIMB) créée par le ministère du développement durable et des transports en 2010. D'où ce travail exploratoire en vue de comprendre comment les idées du web sémantique pourraient s'appliquer à l'information transport.

C. Objectifs de l'étude

Ce travail a été réalisé dans le cadre d'un stage par Julien Plu, étudiant en 1ère année Master Informatique de l'université de Montpellier 2 en juin, juillet et août 2011 suite aux premiers contacts entre le CETE et François Scharffe, du LIRMM[2], le coordinateur du projet Datalift, et grâce au soutien de la DGITM et de l'AFIMB.

L'objectif de ce stage est de démontrer comment publier des données concernant le transport collectif (TC) sur le web sémantique, en suivant la démarche du projet Datalift sur la mise en œuvre du web sémantique, financé par l'ANR en 2010, et auquel participent entre autres l'Insee et l'IGN.

La faisabilité de publier des données TC sur le web des données sera étudiée concrètement à partir de 2 sources de données : l'annuaire des services d'information www.passim.info et les fichiers XML au format Neptune des lignes de transport public des Conseils Généraux de l'Isère et de la Gironde.

D. Démarche

Pour ce stage, les différentes étapes du travail effectué ont été les suivantes :

- Prise de connaissance du contexte : contenu de l'annuaire passim, exemple de fichier XML Trident (Neptune) exporté par le logiciel Chouette, modèle conceptuel de données Transmodel sous-jacent.
- Proposition d'ontologies pour les services d'information Passim, pour l'offre de transport public Neptune sur la base de Transmodel.
- Transformation de la base Passim en RDF ; développement d'un outil de traduction du format Neptune vers RDF.
- Interconnexion avec des jeux de données sur le web de données.
- Description de la chaîne technique (justification du choix des outils mis en œuvre, pérennité et évolutions attendues).
- Perspectives d'utilisation du web de données pour publier l'information transport : proposition d'exemples de requêtes montrant la pertinence d'une publication sur le web de données par rapport à la 'simple' publication de fichiers sur le web ou la mise en place de web services
- Bilan et suites à donner.
- Mémoire de fin de stage, qui a fait l'objet d'une refonte par le LIRMM avec le CETE en vue de sa publication.

E. Contenu du rapport

Outre la présente introduction, ce rapport téléchargeable sur le site web du CETE comprend 3 parties centrales : tout d'abord, la présentation de la chaîne complète et des principaux types d'outils permettant de passer de données structurées à leur publication sur le web sémantique, puis leur mise en pratique dans les deux parties suivantes, décrivant comment nous avons procédé pour publier 2 types de données : d'une part l'annuaire des services d'information Passim, d'autre part la description de l'offre théorique d'un réseau de transport public Neptune.

Il se termine par une conclusion proposant des pistes pour continuer, et en annexe, par une liste des principales références utiles.

[2] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier

II. METHODE ET CHAINE DE PUBLICATION SUR LE WEB DES DONNEES

A. principes et notions techniques

1. Qu'est-ce qu'une ontologie ?

L'ontologie est la base de ce que l'on appelle la représentation des connaissances, un domaine de l'informatique et de l'intelligence artificielle né de la volonté des chercheurs de représenter diverses connaissances humaines sous une forme utilisable par des ordinateurs pour effectuer des raisonnements. Ces connaissances sont exprimées sous formes de symboles (les données) auxquels on donne une signification : une « sémantique ».

Imaginons la problématique suivante : ayant une base de documents (textes, images...) et une requête, comment trouver les documents pertinents pour cette requête ?
En général, votre moteur de recherche préféré recherche la suite de mots de votre requête dans les textes des documents indexés (recherche *full-text*), ou dans des listes de mots associés au document (aussi appelées méta-données).

Par exemple, on tape dans notre moteur de recherche préféré les mots suivants pour rechercher des images : « automobile » puis « voiture ». On s'aperçoit que les résultats ne sont pas du tout les mêmes, alors que, les deux mots représentant la même chose, on pourrait s'attendre à trouver les mêmes images.

Que se passe-t-il ? En fait, le moteur de recherche compare des mots sans prendre en compte leur sémantique. Il exécute une recherche purement syntaxique, sans gestion des synonymes, sans tirer parti du fait que « automobile » et « voiture » représentent le même concept. Plus précisément, on peut dire qu'il n'y a pas de gestion de la relation de généralisation/spécialisation sur les concepts. Par exemple, « taxi » est une spécialisation du concept de « voiture », et « véhicule » est une généralisation du concept de « voiture ». Ainsi, pour raisonner, il ne faut plus se baser sur les mots mais sur les concepts. Le terme 'raisonner' est très souvent employé en intelligence artificielle, il signifie l'action de produire des connaissances à partir de connaissances.

2. Comment décrire des données et des connaissances ?

Ainsi, pour résoudre ce problème, on construit des bases de connaissances, constituées de :

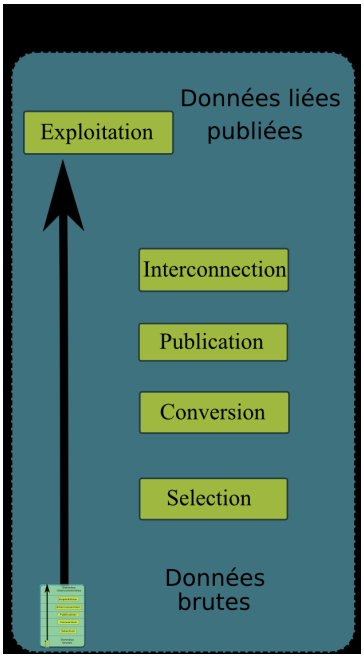
- une **ontologie** : un ensemble de concepts et de relations entre ces concepts ;
- des **règles** : une expression de contraintes sur les relations et concepts de l'ontologie ;
- des **faits** : des 'individus' de l'ontologie.

En pratique, la publication de données sur le web des données revient à rendre accessibles via le web dans un format approprié des faits, décrits par le vocabulaire et les règles d'une ontologie. Ces faits sont décrits en tant que triplets RDF, et publiés sous divers formats (voir plus loin).

B. La Chaîne de publication des données

Comme fil conducteur pour décrire le travail réalisé avec les données de transport public, nous reprenons ici les 5 étapes pour transformer des données brutes en données liées que distingue le projet Datalift : sélection, conversion, publication, interconnexion, et exploitation des données (dans des applications).

« L'ascenseur des données (datalift) » figure extraite du site datalift



1. Sélection (conception ou utilisation d'une ontologie)

Pour pouvoir tirer parti de la sémantique, il faut modéliser l'information ; pour cela, on s'appuie sur une ou des ontologies :- Si les données sont dans des domaines pour lesquels il existe déjà des ontologies bien connues, on les utilise.

- Sinon, on est obligé de définir une nouvelle ontologie, en espérant qu'elle sera réutilisée par d'autres. Comme le web des données est encore en chantier, beaucoup de domaines métier (voire la plupart) ne disposent pas (encore) d'ontologies communément acceptées, alors même que des modèles métier existent par ailleurs pour ces domaines (puisque des logiciels métier et des bases de données existent), dont certains sont publics et même normalisés (exemple de Transmodel pour le Transport Public). La nouvelle ontologie est publiée, le cas échéant. Le standard pour décrire une ontologie est OWL et RDFS. RDF Schema fournit des éléments de base pour la définition d'ontologies ou vocabulaires destinés à structurer des ressources RDF : les composants principaux de RDFS sont intégrés dans un langage d'ontologie plus expressif (mais plus complexe), OWL. (Cf. plus loin le § D : Outils).

2. Conversion

Ensuite, pour rendre un jeu de données particulier utilisable, il faut le publier sous une forme permettant de savoir que les données sont des informations rattachées à des concepts définis dans des ontologies. Un jeu de données se décrit comme une liste de faits sous forme de **triplet** : telle donnée est une information (dans une ontologie existante) qui a telle valeur. Le standard du web sémantique pour cela est RDF (cf. http://fr.wikipedia.org/wiki/Resource_Description_Framework pour des détails et des exemples).

3. Publication

Une fois au format RDF, il faut publier ces données (après s'être assuré que les ontologies auxquelles nos données se réfèrent sont elle-mêmes publiées sur le Web, cf. ci-dessus 1. Sélection).

Pour cela, nous pouvons :

- soit tout simplement publier le fichier RDF sur le Web, c'est à dire rendre le fichier accessible depuis un navigateur par une url (http://...fichier.rdf);
- soit les publier par l'intermédiaire d'un *SPARQL endpoint*. Un SPARQL endpoint permet aux utilisateurs (humains ou autres) d'interroger une base de connaissances via le langage SPARQL (qui est une adaptation du langage de requête sur les bases de données, SQL, aux requêtes sur le web des données).

La publication comprend aussi la production de *méta-données* permettant le référencement dans divers moteurs de recherche du web sémantique, comme Sindice. Pour cela, il y a certaines prérogatives à respecter :

- utiliser un vocabulaire de description de données nommée VoID ; (*Vocabulary of Interlinked Datasets* cf. http://tcuvelier.developpez.com/tutoriels/web-semantique/indexation-donnees-void/)
- créer un *semantic sitemap*, similaire à un fichier sitemap classique d'un site web, mais adapté aux besoins des moteurs de recherche du Web sémantique ;
- créer un package sur le site de la CKAN[3] permettant de référencer les données RDF dans cet annuaire de bases de connaissances ;
- se référencer manuellement dans Sindice ou même Google.

4. Interconnexion

Rappelons que web signifie toile, pour indiquer qu'il s'agit de relier entre eux des documents. De même que si, lorsqu'on publie une page HTML en la rendant accessible sur un serveur HTTP via une URL) elle n'est pas encore reliée au Web (il faut pour cela qu'en outre son adresse soit référencée depuis d'autres pages HTML), il ne suffit pas que le fichier RDF contenant les triplets décrivant le jeu de données soit publié sur le Web (c'est-à-dire qu'on puisse y accéder depuis un navigateur en appelant son URL) pour que le fichier soit considéré comme faisant partie du Web des données. Il faut encore relier les données à d'autres données déjà présentes sur le Web (des données) : soit les référencer sur les principaux moteurs de recherche, soit les référencer à partir de pages existant déjà sur le Web (soit les deux, évidemment).

Pour relier nos données au web sémantique, donc, il faut :

- soit procéder à un référencement classique sur un moteur de recherche pour le Web de données ; (exemple http://sindice.com/main/submit)
- soit déterminer par nous-même (ou à l'aide d'un outil) des URI équivalentes entre nos données et celles situées sur d'autres sources de données.

5. Exploitation (Utilisation des données liées dans des applications)

Une fois nos données brutes transformées et publiées sur le web des données, en suivant la chaîne décrite ci-dessus, il s'agit désormais de les exploiter dans des applications possibles, qui sont in fine la raison d'être du web sémantique. Nous donnons ici quelques éléments généraux sur les applications possibles de ces données, car il est difficile de donner des exemples d'utilisation réelle, pour le 'grand public'.

a) Les données liées sont des données libres :

Le point essentiel est que les applications possibles sont limitées aux données publiées. Le succès du web sémantique dépend donc dans une large mesure de la généralisation de l'open data. A priori les producteurs de données publiques réutilisables pour le web sémantique peuvent être des administrations et services publics, des particuliers et associations, des laboratoires scientifiques, et éventuellement des entreprises privées.

Le web sémantique implique des données publiques (au sens de leur libre réutilisation), c'est la suite 'logique' de l'open data. Les démarches open data des administrations nationales et locales ont un objectif essentiel : favoriser la réutilisation des données publiques. Une démarche similaire existe avec les données scientifiques (open science).

L'open data va inciter à standardiser les données d'un même métier ou domaine ; en effet, la standardisation facilite grandement la réutilisation, dans le sens qu'une fois qu'une application est

[3] CKAN est un logiciel développé et maintenu par l'Open Knowledge Foundation, association à but non lucratif promouvant l'open knowledge, y compris l'open content et l'open data, et une plate-forme (utilisant ce logiciel) CKAN.net dont il est question ici pour référencer des données sur le web sémantique.

développée avec un jeu de données (par exemple les horaires de bus pour une ville), elle pourra fonctionner 'pour le même prix' avec d'autres données.

Le web sémantique est l'étape suivante, après une éventuelle généralisation de la publication de données librement réutilisables. L'open data permettra le développement d'applications ciblées sur certains domaines et leur généralisation très rapide si les données sont publiées de manière standard. La promesse du web sémantique est de permettre la réutilisation de données par des utilisateurs qui ne connaissent pas a priori un domaine métier particulier.

b) Peu d'applications 'visibles' du web sémantique mais de larges perspectives d'applications :

Aujourd'hui, les principales applications du web des données visent surtout à :

- Rechercher directement des données sur le web sémantique (pour cela il existe SPARQL, RQL et RDQL) ;
- Améliorer la recherche et le traitement des données du web 'classique' en le 'sémantisant'.

Pour l'instant, le Web des données est essentiellement produit par des chercheurs :

- il y a beaucoup de logiciels et d'outils, mais la plupart d'entre eux n'en sont pas encore à une version stable ;

- les premières données utilisées sont les données libres produites par des initiatives comme wikipedia ou openstreetmap ; plus récemment, le mouvement open data dans les administrations permet d'élargir le champ des données réutilisables.

Jusqu'à récemment, les industriels du web (Google, Amazon, Oracle ou autres Microsoft) ne produisaient quasiment pas d'outils pour ce domaine, mais cela commence heureusement à changer depuis environ 2 ans. L'avenir sera sans doute un mélange des deux: évolution des grands groupes et start-ups qui auront réussi.

L'avènement du web sémantique est encore incertain, il est impossible d'annoncer une date prochaine où le web des données sera devenu une réalité utilisée par le grand public, ni sous quelle forme cela se concrétisera. Mais une chose est sûre, le web sémantique prend de plus en plus d'importance au fur et à mesure des années. Actuellement, dans le grand public, on parle beaucoup de Cloud mais cela n'a strictement rien à voir avec le Web sémantique : le Cloud est juste un moyen technique d'héberger ses données, applications et les infrastructures de son entreprise dans le Web et ainsi confier cela aux industriels (par exemple Amazon, Google, Microsoft avec Azure, etc.).

c) Difficultés de la réutilisation des données :

La réutilisation des données liées implique des difficultés de principe qu'il est important d'avoir à l'esprit. Pour commencer, comme dans le monde réel, l'utilisateur humain a une part d'interprétation dans son utilisation de la sémantique. C'est le problème de la polysémie : une même donnée peut être décrite par plusieurs ontologies (par exemple, selon le contexte, une voiture est à la fois un moyen de transport, un patrimoine, un produit industriel).

Comme dans le monde réel aussi (et comme sur le web des documents), toutes les données publiées n'ont pas la même valeur (de même que tous les documents n'ont pas la même valeur). Les données de mesure de laboratoires scientifiques ou les données statistiques d'administrations devraient pouvoir être utilisées avec un bon niveau de confiance ; néanmoins, même ces données ne peuvent évidemment pas être considérées comme des vérités absolues :

- les données dépendent du contexte dans lequel elles ont été produites ;

- la définition de telle ou telle grandeur (dans une ontologie) n'est pas forcément identique d'un jeu de données à l'autre

- il y a bien sûr des erreurs dans les données (intentionnelles ou non).

Typiquement, le contenu de Wikipédia est aujourd'hui une des principales sources de contenu pour le web des données (pour la simple raison qu'il s'agit d'une des principales sources de données assez structurées et générales librement réutilisables). On sait bien pourtant qu'il y a des erreurs dans Wikipédia.

Au total, il y a donc forcément des contradictions sur le web des données, que les outils et applications ne savent pas traiter (ce n'est d'ailleurs pas le rôle de l'application finale de vérifier la validité des données). De la même manière qu'il y a des pages web qui disent des choses contradictoires, il y a des fichiers RDF qui énoncent des faits contradictoires, ou ambigus (par exemple, plusieurs villes s'appellent Paris en France, ou aux USA et ailleurs).

C. Outils

Après avoir présenté dans la section précédente la chaîne des données de publication de données liées, nous présentons ici, pour chaque étape de la chaîne, les outils et des langages utilisés pour notre projet :

Sélection :

- OWL, pour le langage d'écriture de l'ontologie ;
- Neologism, pour la publication et la documentation de l'ontologie.

Conversion / Interconnexion :

- Google Refine, outil de nettoyage de données.

Publication :

- Sesame, pour stocker ces données sous forme de triplets ;
- PostgreSQL, pour l'hébergement de la base de données ;
- Apache, pour le serveur web afin d'avoir les données toujours accessibles sur le Web ;
- Apache Tomcat, pour le serveur d'application, hébergeant D2R et Sesame.

Exploitation :

- Jena[4], pour le framework de manipulation de données RDF et de requêtes SPARQL.

1. Sélection

a) OWL

OWL (ou *Ontology Web Language*) est une recommandation émanant du W3C permettant de définir et d'instancier des ontologies. Ce langage constitue une extension des langages RDF et RDFS et permet de combler leur manque d'expressivité. OWL 'introduit notamment des notions de classes ou de propriétés équivalentes, d'égalités entre instances, de propriétés symétriques, de restrictions de valeurs...

OWL comprend trois sous langages :

- OWL Lite, tout d'abord, est le moins expressif des trois. Ce langage est particulièrement adapté aux personnes souhaitant bénéficier d'une expressivité plus importante que RDF/RDFS tout en conservant une certaine simplicité d'utilisation ;
- OWL DL est un sous-langage offrant une expressivité maximale. Toutes les propriétés OWL sont ainsi présentes dans ce langage. Toutefois, un ensemble de contraintes a été fixé afin de garantir deux propriétés :
  - OWL DL peut résoudre l'ensemble des problèmes d'inférences (complétude) ;
  - OWL DL assure que ces problèmes peuvent être résolus en un temps fini (décidabilité).
- OWL Full, un sous-langage permettant d'avoir une expressivité maximale et une grande liberté dans la conception des ontologies. OWL Full lève la plupart des verrous fixés dans OWL DL. Cette liberté syntaxique a toutefois un prix : il est impossible de garantir que les problèmes d'inférences concernant une ontologie utilisant OWL Full pourront être résolus

[4] cf. http://j***ena***.*sourceforge.net* Jena était envisagé pour ce projet mais n'a pas été mis en oeuvre en pratique, il n'est donc pas décrit ici.

en un temps fini. Parmi les libertés offertes par OWL Full, on trouve notamment la possibilité d'utiliser une classe comme une instance ou encore la possibilité d'intégrer plus facilement des éléments de RDF/RDFS dans l'ontologie.

Le langage que nous avons utilisé pour les jeux de données Passim et Neptune est OWL en version Full.

b) Neologism

Neologism est un logiciel libre développé par le DERI (laboratoire d'informatique de l'université de Dublin) que le LIRMM a installé sur un de ses serveurs : Neologism permet d'éditer et publier une ontologie.

Neologism est basé sur le système de publication de contenu web (CMS) Drupal, il est donc très simple à installer. Neologism produit un site Web qui peut notamment créer une ontologie, en publier ou en importer une, et la modifier ainsi que la documenter. Comme tout se passe en ligne, il est possible de faire toutes ces actions de n'importe où.

2. Conversion des données

Ensuite, une fois l'ontologie décrite en OWL et publiée, il faut pouvoir transformer les données en format RDF afin de les publier elles aussi. Pour cela, nous nous sommes appuyés sur l'outil Google Refine.

Google Refine est la nouvelle mouture de Freebase Gridwork acquis par Google suite au rachat de Metaweb / Freebase. Cet outil permet de charger de grandes quantités de donnés, et d'effectuer de manière automatisée des opérations qui manuellement pourraient être très lourdes (comme la fusion de cellules, extraction de données, vérification et correction des données, ajout automatisé de colonnes, etc.). D'autres logiciels similaires existent, comme le logiciel libre Talend, mais Google Refine se distingue d'autres outils par une interconnexion plus naturelle avec le web sémantique : il est possible de se connecter à des services web pour transformer ou récupérer de nouvelles données depuis internet (notamment Freebase et DBPedia). Par exemple, il est possible de géocoder des adresses d'un fichier chargé grâce à l'API de Google Maps : Google Refine traite les données XML retournées par Maps et peut aussi traiter des réponses au format JSON. Une fois les données traitées, on peut les transformer au format de son choix, en l'occurrence pour nous : RDF.

Dans la documentation de Refine, il est fait mention que l'on pourrait charger des fichiers Excel mais il semble que cela ne fonctionne pas parfaitement, nous sommes donc partis du contenu de l'annuaire Passim au format texte (CSV). Il est bien sûr possible de transformer les données manuellement mais aussi de définir des actions plus sophistiquées en utilisant des conditions et des règles portant sur des données Texte, Tableau, Numérique ou Date, avec la possibilité d'utiliser des conditions (if, or, and..) ou des boucles.

3. Publication

a) Quel format de publication des données ?

De nombreuses technologies sont associées au Web sémantique. Parmi les plus connues, on peut citer RDF (Ressource Description Framework) qui modélise l'information simplement sous forme de ressources décrites par des triplets et permet l'échange de données sous divers formats pour communiquer entre différentes applications (RDF/XML, RDF/JSON, N3, Turtle, N-Triples et d'autres). Dans le domaine du Web sémantique, la sémantique des données est décrite par des ontologies– décrivant concepts, termes ou relations avec des langages dont les principaux sont RDFS (*Resource Description Framework Schema*) et OWL (*Web Ontology Language*). Il existe aussi des langages de description des données structurées dans du XHTML afin que des outils effectuent un traitement automatique de ces différentes données. Ces langages sont RDFa et Microformat. Ensuite, pour finir avec la liste des technologies, il existe un langage de requête, au même titre que SQL pour les bases de données relationnelles, SPARQL, qui effectue des requêtes sur des triplets RDF. Il en existe d'autres (RQL et RDQL), mais ils sont bien moins utilisés.

Depuis peu, les principaux éditeurs de moteurs de recherches (Microsoft, Yahoo et Google) travaillent sur une norme pour justement sémantiser les pages Web autrement qu'avec du RDFa ou du Microformat. Ils tirent parti d'un nouveau système de sémantisation des données implémenté dans le HTML5 appelé Microdata. Leur norme est consultable sur le site schema.org. Précisons que les annotations dans les pages HTML ne servent principalement qu'aux moteurs de recherche et rarement aux applications, c'est pour cette raison que, de manière complémentaire, sont publiées des bases de données RDF.

b) Outil de publication RDF

Sesame[5] est un framework Java open source pour interroger et stocker des données RDF. Il était à l'origine développé par l'entreprise allemande Aduna comme un prototype de recherche pour le projet de recherche de l'union européenne « On-To-Knowledge ». Il est maintenant développé comme un projet commun et hébergé sur openrdf.org. Sesame a une excellente interface d'administration qui est inclus dedans, il est facile à installer et il a de très bonnes performances. Sesame s'utilise depuis un navigateur : une fois les données nettoyées et formatées, il faudra les exporter en RDF et les inclure dans le triplet store Sesame (: outil pas présenté pour l'instant ?), ce qui permet par la suite de pouvoir interroger ces données avec des requêtes SPARQL.

D. Le web sémantique pour les transports publics

Cette dernière section du chapitre II fait le lien avec l'application des outils de la chaîne des données aux données de transport public dans les 2 chapitres précédents.

1. recensement de l'existant

Très grossièrement, on peut dire que le web sémantique est d'origine européenne et académique, avec le Royaume-Uni et l'Allemagne comme pays les plus avancés, le Royaume-Uni présentant l'avantage en outre d'être comme les USA dans une culture anglo-saxonne de publication des données publiques et donc d'open data ; et les Etats-Unis, qui commencent à s'y intéresser concrètement à partir du moment où il peut y avoir des retombées économiques et des créations d'entreprises.

On retrouve cette situation pour le domaine du transport public. Le Royaume-Uni est le plus avancé pour plusieurs raisons :

- ce sont les pionniers de l'open data en Europe (culture de transparence publique, Crown Copyright, etc.) ;
- le domaine du transport public a bénéficié depuis une dizaine d'années d'un fort investissement dans la normalisation des données (projet Transport Direct, normes Naptan, Nptg, TransXchange etc. : http://www.dft.gov.uk/public-transportdatastandards/) ;
- à l'origine du web sémantique (Tim Berners-Lee).

Un très bon exemple de ce qu'il est possible de faire est publié sur le site http://data.gov.uk . Beaucoup de données de transport public au Royaume-Uni sont sur le web, bien spécifiées, mais peu sont encore des données liées (c-à-d au bon format pour le web des données). Seuls deux jeux de données conformes aux standards du web sémantique ont été identifiés :

- le référentiel national des arrêts de bus du Royaume-Uni (NAPTAN) ;
- les horaires de bus de Manchester (http://ckan.net/package/greater-manchester-bus-timetable-linked-data).

Malgré le grand nombre de réseaux de TC américains dont l'offre est publiées en open data, leur traduction en données liées n'a fait que l'objet à notre connaissance que de quelques tentatives et nous a semblé bien moins avancée qu'au Royaume-Uni.

[5] http://www.aduna-software.com/technology/sesame

D'autres données devraient suivre, notamment l'ensemble des horaires des TC du Royaume-Uni en 2012 (National Public Transport Data Repository).

En France, on observe depuis 1 an ou 2 une intérêt fort pour la publication de données 'open data' dans les transports publics. Pour l'instant (à l'été 2011), seuls 3 réseaux sont en ligne à notre connaissance : Rennes, Bordeaux, Gironde, mais beaucoup sont en préparation et seront ouverts dans les mois qui viennent. A partir de ces données publiées à des formats standard (Neptune, GTFS le standard Google), le travail technique nécessaire pour assurer leur publication en tant que données liées sur le web sémantique ne devrait pas être trop important, une fois la chaîne de publication bien définie, comme la suite de ce rapport veut le démontrer.

2. Perspective d'applications

Les données de transport public peuvent être utilisées dans d'innombrables applications professionnelles ou grand public et commencent à l'être à mesure que les données sont publiées et réutilisables, ou que des web services publics d'accès à l'information sont mis en ligne. Par exemple, une carte en temps réel de tous les véhicules (trains, métros, bus, trams, taxis) d'une ville donne à l'usager une visualisation d'ensemble et des possibles retards: cette application existe (trains suisses, bus anglais...), sans qu'il soit besoin pour cela du web des donnés.

L'apport du web sémantique par rapport aux technologies du Web classique, est de pouvoir développer des applications sans avoir besoin de connaissances particulières des données métier du TC. Par exemple, il sera possible en utilisant des outils génériques et des requêtes bien formulées, sans développement logiciel spécifique, d'afficher les restaurants et autres activités possibles autour de chaque arrêt de transport (avec, si c'est un cinéma, les horaires des films, si c'est un restaurant : les différents menus et spécialités, etc.), ou de trouver des parcours touristiques en TC en fonction du lieu de destination et des centres d'intérêt de l'utilisateur.

Le web sémantique doit aussi permettre sans aucun développement logiciel, par exemple, de savoir combien d'arrêts TC en France s'appellent 'Victor Hugo' ou combien d'arrêt portent le nom d'un écrivain français, et parmi ces arrêts, lesquels sont situés dans une voie qui porte le même nom, ou de connaître le nombre de lignes de bus dans une ville en reliant la connaissance de l'offre TC à des sources telles que DbPedia et GeoName. Ces exemples d'application n'ont certes pas en soi un intérêt phénoménal en termes de retombées économiques, mais comme dans le même esprit, n'importe quelle autre requête portant sur 2 domaines ou plus pourrait être formulée (transport et logement, transport et population, transport et info culturelles ou touristiques, etc.), au total le web sémantique promet d'apporter une réelle valeur ajoutée par rapport à de 'simples' plates-formes open data.

III. PUBLICATION DE SERVICES D'INFORMATION TRANSPORT (PASSIM)

A. L'annuaire passim

L'annuaire PASSIM[6] est édité par le CERTU (Centre d'Etudes sur les Réseaux, les Transports et l'Urbanisme) ; il recense et met à disposition une liste de services d'information français sur les transports de voyageurs et autres services de mobilité. Son contenu est géré par le CETE Méditerranée.

En pratique, l'annuaire est un site web (fixe et mobile) qui permet de retrouver les services pertinents pour une commune ou un territoire en France, en distinguant les modes ou types de services de transport (voiture, transport collectif, etc.) et le périmètre (urbain, départemental, régional). Les services sont au moins des sites web, parfois des services téléphoniques, ou des applications pour mobiles (à l'avenir ce pourrait aussi être des web services).

L'annuaire contient des données librement réutilisables (qui ont vocation à être dans data.gouv.fr, la plate-forme open data de la mission Etalab). Cet annuaire pourrait être étendu par exemple pour référencer les données open data ou les web services publics permettant de récupérer des informations TC. Des idées similaires d'annuaire des services d'information transport existent au niveau de la Commission Européenne.

B. les étapes pour les données Passim

En reprenant les étapes de la chaîne des données décrites plus haut, voici celles que nous avons mises en oeuvre pour les données de l'annuaire Passim, que nous décrirons dans la suite de cette partie.

Sélection :

- Étape 1 : conception de l'ontologie ;
- Étape 2 : développement de l'ontologie ;
- Étape 3 : publication de l'ontologie.

Conversion / Interconnexion :

- Étape 4 : utilisation de Google Refine pour la transformation des données en RDF et l'interconnexion ;

Publication

- Étape 5 : publication des données RDF dans Sesame.

C. Ontologie

Nous n'avons pas trouvé d'ontologie générale réutilisable décrivant un annuaire de services d'information, qui aurait pu être spécialisée au domaine de Passim : l'information transport.
Nous avons donc décidé de construire notre propre ontologie pour Passim, en la reliant néanmoins autant que possible à des vocabulaires existants. Par exemple, nous avons utilisé l'ontologie de l'INSEE pour les départements et les régions, issue des travaux de Datalift.

L'ontologie de Passim contient 4 classes et 18 propriétés. Elle est publiée sur : http://data.lirmm.fr/ontologies/passim

[6] www.passim.info

Elle contient des classes :

passim:TransportServiceInformation	Cette classe représente un service d'information de transport.
passim:Mode	Cette classe représente les différents modes de transports couverts par le service d'info
passim:Service	Cette classe représente les services d'info transports.
passim:Coverage	Cette classe représente la couverture géographique d'un service d'info.

... et des propriétés :

passim:SMSInformation	indique si le service est accessible par SMS
passim:centerTown	ville principale couverte par le service
passim:comment	commentaires relatifs au service d'info
passim:department	Départements couverts par le service.
passim:infoPoint	adresse du point d'information (guichet).
passim:isAccessibilityForDisabledPerson	indique si le service d'info transport comprend des informations pour les personnes handicapées ou non. (info sur l'accessibilité du transport)
passim:isWebSiteAccessibilityForDisabledPerson	indique si le site Web du transport est accessible aux handicapés (accessibilité de l'info)
passim:landInformation	commentaire décrivant les informations sur le terrain (afficheurs, etc.)
passim:cityThrough	liste des communes couvertes par le service
passim:mobileApplication	adresse de l'application mobile
passim:modeOfTransport	liste des modes de transport couverts par le service d'info
passim:postalCode	code postal de la ville principale du transport.
passim:region	Région administrative couverte par le service
passim:remark	texte de remarques concernant le service d'info
passim:serviceCoverage	couverture géographique du service d'info
passim:serviceName	nom du service.
passim:typeOfService	type de service : liste parmi horaires, perturbations en temps réel, recherche d'itinéraires, etc
passim:webSite	adresse du site Web du service d'info

D. Conversion

Les données exportées en CSV depuis le site web de Passim ont été converties en RDF avec l'outil Google Refine.
Une erreur dans les données du fichier CSV a fait que certains services d'info référencés dans passim n'ont pas pu être convertis correctement en RDF. L'erreur est en fin de fichier RDF (la balise <LandInformation> contient toutes les données qui n'ont pu être traitées). Les données converties et publiées en RDF couvrent néanmoins la majorité du contenu de l'annuaire passim et permettent de se faire une idée

Passim a été référencé sur Sindice, ce qui en principe permet de faire des requêtes incluant des données de Passim depuis ce moteur de recherche (néanmoins en pratique cela ne semble pas avoir fonctionné).

Nous avons enfin publié un 'package' passim sur le site de la CKAN, qui peut être trouvé à cette adresse : http://ckan.net/package/passim.

E. Publication

Les données de Passim ont été rendues disponibles sur ce SPARQL endpoint : http://data.lirmm.fr/openrdf-workbench/repositories/Passim/query

On peut ainsi faire des requêtes sur ces données. Comme par exemple, la liste des villes desservies par les lignes de la compagnie TaM (qui est le réseau de TC urbain de Montpellier) :

```
SELECT DISTINCT ?ville WHERE {
    ?s passim:serviceName ?o .
    ?s passim:cityThrough ?ville
    FILTER (?o = "TaM")
}
```

F. Un premier bilan

Malgré quelques problèmes techniques de conversion avec Refine et de publication (par exemple sur Sindice), on peut affirmer que le contenu de passim est publiable sans difficulté majeure et pourrait faire l'objet d'applications pour le Web sémantique, par exemple lorsqu'il sera publié sur Etalab. L'ontologie pourra être revue à cette occasion de manière à relier le jeu de données de Passim à plus de données pertinentes du web sémantique.

IV. DESCRIPTION DE L'OFFRE DE TRANSPORT (NEPTUNE)

A. Profil Neptune

NEPTUNE signifie « Norme d'Echange Profil Transport collectif utilisant la Normalisation Européenne ». Issue du projet Européen TRIDENT puis de travaux français relatifs à l'application CHOUETTE, NEPTUNE est une norme française (NFP-99506) spécifiant le format de référence pour l'échange de données d'offre théorique TC, particulièrement utile pour la mise en place de systèmes d'information multimodale. Les spécifications NEPTUNE se composent d'une part d'un modèle conceptuel de données en UML (issu du projet TRIDENT, basé sur Transmodel V4.1) relatif à la définition du réseau (lignes, arrêts) et du service théorique (courses, horaires), d'autre part d'un schéma XSD. Les données échangées se présentent sous la forme d'un répertoire comprenant un fichier XML par ligne, chaque fichier décrivant l'ensemble des informations relatives à une ligne (arrêts, horaires, etc.) .Le profil NEPTUNE est totalement compatible avec le logiciel CHOUETTE dont le développement est soutenu par le ministère des transports (cf. http://www.chouette.mobi/spip.php?rubrique40). Neptune évoluera dans le cadre des travaux de normalisation européens Netex.

Il est à noter que Google a spécifié un format d'échange au format texte, GTFS, qui est le standard pour la publication de données TC en open data (essentiellement par des réseaux nord-américains).

Les premières données Neptune open data ont été publiées par le Conseil Général de la Gironde, d'autres suivront (le Conseil Général de l'Isère, par exemple, a bien voulu mettre à disposition ses données Transisère dans le cadre de nos tests ; la Saône et Loire par exemple se prépare à l'open data).

B. les étapes pour les données Neptune

En reprenant les étapes de la chaîne des données décrites plus haut, voici celles que nous avons mises en oeuvre pour les données Neptune, que nous décrirons dans la suite de cette partie.

Sélection :

- Étape 1 : conception de l'ontologie ;
- Étape 2 : développement de l'ontologie ;
- Étape 3 : publication de l'ontologie.

Conversion :

- Étape 4 : développement d'un outil de transformation des données Trident en RDF ;

Interconnexion :

- au niveau du code que nous avons écrit pour convertir des données Neptune en RDF, l'interconnexion se fera avec Geonames par l'intermédiaire des coordonnées GPS.

Publication :

- Étape 5 : publication des données RDF dans Sesame.

C. Ontologie

L'état de l'art des ontologies existantes montre que deux ontologies sont éventuellement réutilisables

http://transport.data.gov.uk/doc/bus-stop-point

http://ckan.net/package/greater-manchester-bus-timetable-linked-data

Les formats de donnés anglais (schéma XML TransXchange) sont 'cousins' du format Neptune, il y aura sans doute a minima des liens à faire entre vocabulaires.

A terme (à partir de fin 2012 au mieux), le standard européen Netex (en cours de développement) devrait (définitivement espérons-le) unifier le vocabulaire et les formats d'échange de données en Europe.

A court terme néanmoins, on ne peut pas relier les données Neptune au reste du web des données sauf via les coordonnées géographiques des arrêts ; notre vocabulaire semble isoler des vocabulaires existants. En outre, pour simplifier les développements dans le cadre de ce stage, nous nous sommes limités à la topologie du réseau (lignes, arrêts), et nous n'avons pas décrit les horaires dans notre ontologie.

L'ontologie de Neptune contient 4 classes et 10 propriétés. Détaillons maintenant cette ontologie. Cette ontologie est accessible à : http://data.lirmm.fr/ontologies/neptune

L'ontologie comporte 4 classes :

neptune:Company	Cette classe représente l'entreprise exploitant le réseau de transport public. C'est une sous-classe d'une des classes de l'ontologie FOAF, Organization.
neptune:Line	Cette classe représente une ligne du réseau de transport.
neptune:PTNetwork	Cette classe représente un réseau de transport public.
neptune:StopPoint	Cette classe représente les arrêts d'une ligne.

... ainsi que des propriétés :

neptune:companyName	nom de l'entreprise exploitant le réseau
neptune:containLine	ligne faisant partie du réseau.
neptune:containStop	arrêt de transport public
neptune:leadsNetwork	réseau de transport public
neptune:networkName	nom du réseau.
neptune:number	numéro de ligne.
neptune:stopComment	commentaire concernant un arrêt.
neptune:stopLatitude	latitude d'un point d'arrêt. C'est une sous-propriété d'une des propriétés de l'ontologie GEO du W3C, lat.
neptune:stopLongitude	longitude d'un point d'arrêt. C'est une sous-propriété d'une des propriétés de l'ontologie GEO du W3C, lon.
neptune:stopName	nom d'un point d'arrêt

D. Développement de l'outil de traduction

Pour les données Neptune, disponibles directement en XML, l'outil Google Refine n'est pas approprié : il est nécessaire de développer un programme spécifiquement chargé de la conversion en RDF.

Nous avons développé cet outil en Java. Il s'occupe de traduire toutes les données correspondantes à l'ontologie contenues dans les fichiers Neptune en triplet RDF, plus précisément dans le format N-TRIPLES. Les URI correspondant aux ressources RDF s'écrivent de la manière suivante :

http://data.lirmm.fr/neptune/xxx_id

où le terme « xxx » représente le nom du transporteur, d'un réseau ou d'un arrêt ou bien le numéro

d'une ligne décrit dans le fichier Neptune, et le terme « id » désigne l'identifiant de ces items. Cela suppose que les identifiants des objets sont bien uniques et pérennes (ce qui n'est pas forcément le cas aujourd'hui, malheureusement, en tout cas au niveau national).

Le code source est disponible sur un dépôt SVN :
http://subversion.assembla.com/svn/neptunetordf/.

E. Démonstration

Pour utiliser cette application de conversion de Neptune en RDF, il suffit de taper la ligne de commande « `java –jar NeptuneToRDF.jar nomdefichier` » où nomdefichier est soit un fichier Neptune soit une archive ZIP.

Nous avons ainsi converti en RDF plusieurs fichiers Neptune du réseau Transisère, puis avons publié le contenu RDF avec l'outil Sesame à l'adresse :

http://data.lirmm.fr/openrdf-workbench/repositories/Neptune/query

Par exemple, la requête suivante demande la liste des noms des arrêts de la ligne 1000 :

```
SELECT DISTINCT ?arret WHERE {
    ?s neptune:number ?o .
    ?s neptune:containStop ?stop .
    ?stop neptune:stopName ?arret
    FILTER (?o = "1000")
}
```

F. Premier bilan

Ce travail a montré qu'une chaîne de conversion et publication automatique de données d'offre TC au format Neptune vers le web sémantique est faisable avec des outils relativement simples. Pour aller plus loin, il faudrait sans doute d'une part étendre l'ontologie à d'autres éléments de données que les lignes et arrêts, et cela aurait certainement intérêt à être fait au niveau européen, d'autre part, référencer systématiquement les données publiées sur les outils du web des données afin de faciliter leur réutilisation dans des applications.

V. CONCLUSIONS ET SUITES A DONNER

Ce travail exploratoire aura permis :

- de présenter les concepts autour du web des données et l'intérêt potentiel pour la réutilisation de données relatives aux TC ;

- d'identifier deux sources de données potentiellement pertinentes pour faire émerger le web des données de transport public ;

- montrer la faisabilité de mise en oeuvre d'une chaîne de publication complète.

Nous n'avons malheureusement pas pu aller jusqu'à produire une démonstration d'utilisation de ces données pour une application concrète, à la fois par manque de temps, et par manque d'une 'masse critique' de données auxquelles se relier. On peut néanmoins se faire une idée des applications en observant ce que font nos collègues anglais, qui sont les plus avancés en la matière.

Toutefois, compte tenu de la dynamique actuelle en matière d'open data et de réutilisation des données publiques y compris dans le transport, qui sont des conditions nécessaires pour le web sémantique, et de la prochaine ouverture de data.gouv.fr qui permettra a minima de publier le contenu de passim, il nous semble que ce premier travail peut être poursuivi dans plusieurs directions :

- à court terme, nettoyer les erreurs résiduelles dans le fichier RDF de passim ;
- envisager dans le cadre d'une initiative open data d'une collectivité locale AOT qui inclurait un volet 'web sémantique', une publication de données d'offre TC Neptune ;
- par la même occasion, développer des applications significatives qui illustrent les utilisations possibles de ces technologies, ce que nous n'avons pas pu faire lors de ce stage trop court ;
- à moyen terme, étendre l'ontologie Neptune à d'autres éléments de données que les lignes et arrêts, et sans doute plutôt au niveau européen en cohérence avec Netex ;
- faire le lien avec les projets nationaux de l'AFIMB (référentiels régionaux de points d'arrêts, standardisation des Web services).

VI. ANNEXE : REFERENCES

Les résultats de cette étude sont disponibles sur notre site web: http://www.cete-mediterranee.fr/tt13/www/

A. Contexte

www.predim.org
www.cete-mediterranee.fr/tt13
www.passim.info
www.chouette.info
http://datalift.org
www.lirmm.fr
http://rdf.insee.fr/geo/

B. Glossaire, termes utilisés

Les définitions se retrouvent facilement notamment via wikipedia.

web des données = web sémantique
linked data = données liées = données publiées aux standards du web des données
URL, URI
Ontologie, règles, faits
OWL (Ontology Web Language)
RDF (Resource Description Framework : RDF/XML, RDF/JSON, N3, Turtle, N-Triples et d'autres), RDFS (RDF Schema)
RDFa Microformat
schema.org
méta-données : VoID (Vocabulary of Interlinked Datasets cf. http://tcuvelier.developpez.com/tutoriels/web-semantique/indexation-donnees-void/)
lod cloud
semantic sitemap
SPARQL endpoint

C. outils :

Neologism logiciel libre développé par le DERI (laboratoire d'informatique de l'université de Dublin) qui permet d'éditer et publier une ontologie.

Google Refine.

Sesame framework Java open source pour interroger et stocker des données RDF

http://www.aduna-software.com/technology/sesame.

Sindice.com
CKAN.net
dbpedia.org

D. Bibliographie

le livre en ligne sur le web sémantique http://linkeddatabook.com/editions/1.0/

Open data :
http://blog.etalab.gouv.fr/
http://www.proximamobile.fr/sites/default/files/RapportDonneesPubliques2011.pdf Rapport de stage sur le web des données (p73) et qui propose 16 recommandations intéressantes dont une sur le

web sémantique (la participation de l'Insee à datalift est citée).
Référentiel de bonnes pratiques pour une plate-forme open data :
https://checklists.opquast.com/opendata/workshop/status/2777

E. Open data & transport public

- USA :
  - http://www.citygoround.org/
- Royaume-Uni :
  - http://data.gov.uk/linked-data
  - http://data.gov.uk/dataset/naptan
  - http://transport.data.gov.uk/doc/bus-stop-point
  - http://data.gov.uk/dataset/nptdr
  - Offre théorique TC (2 mises à jour annuelles) : http://data.gov.uk/dataset/nptdr
  - Londres http://data.london.gov.uk/datastore/package/tfl-timetable-listings
  - Rapport de recherche proposant un cadre général pour les données de Transport : http://www.escience.cam.ac.uk/projects/transport/ntdffinalreport.pdf
  - linked data à Manchester : http://ckan.net/package/greater-manchester-bus-timetable-linked-data
  - SPARQL endpoint pour les arrêts de bus (naptan) : http://openuplabs.tso.co.uk/sparql/gov-transport

- France :

annuaire open data :
http://www.publicdata.eu/package?extras_eu_country=FR&groups=transport
Bordeaux : http://data.lacub.fr/data.php?themes=10
Rennes : http://data.keolis-rennes.com/fr/les-donnees/donnees-telechargeables.html
Gironde : http://www.datalocale.fr/dataset/liste-ligneregulieres-transgironde
www.itinisere.fr données Neptune utilisées du CG38 pour le stage

Formats et normes
Essais d'ontologie TC (peu adaptés à nos besoins)
http://vocab.org/transit/terms/.html
autre : https://github.com/kasei/gtfs-rdf
www.chouette.mobi/normes
Modèle conceptuel de données Transmodel
Neptune (XML Trident/Chouette) : www.chouette.mobi
GTFS (Google) vs TransXchange (ministère des transports britannique)
http://www.dft.gov.uk/transmodel/schema/doc/GoogleTransit/TransmodelForGoogle-09.pdf
et http://www.dft.gov.uk/transxchange/gtfs.htm