Université du Québec à Montréal

Service de la recherche et de la création

LE WEB SÉMANTIQUE ET LES DONNÉES OUVERTES

COMME LEVIERS D'IMPACT DE LA RECHERCHE :

DES COMPÉTENCES FUTURES EN INTELLIGENCE ARTIFICIELLE

Michel Héon PhD, développement et formation

Rachid Belkouch, gestion de projet

1er semestre 2020

Future Skills Centre | Centre des Compétences futures

RESEARCH IMPACT CANADA | Turning research into action

Table des matières

Table des illustrations

1 - Introduction

Ce projet porte sur le développement d'un outil de formation en Web sémantique pour du personnel en gestion des données de la recherche, sous la forme d'un environnement d'apprentissage et d'un guide de formation.

L'avènement des données ouvertes en recherche encourage les universités à améliorer la visibilité, la valorisation et la dissémination de la recherche – soit l'impact de la recherche. Les technologies du Web sémantique permettent de lier des données ouvertes à travers des ontologies. Cette possibilité prend toute son importance dans le monde de la recherche, aujourd'hui où la collaboration est devenue si importante, que ce soit entre les chercheurs eux-mêmes, ou entre les établissements de recherche et l'industrie.

Le Web sémantique est aussi une technologie d'Intelligence Artificielle (IA) qui a fait ses preuves. L'IA est sans aucun doute une compétence professionnelle dont nous aurons besoin dans le futur, car il y a un manque important de ressources humaines au Canada en Web sémantique.

Grâce au financement du Réseau Impact Recherche Canada (RIC)[1] et du Centre de Compétences Futures (FSC)[2], l'Université du Québec à Montréal (UQAM)[3] a développé un environnement d'apprentissage, et réalisé une formation pour une cohorte de 22 professionnels, au cours du premier semestre 2020.

[1] http://researchimpact.ca/
[2] https://fsc-ccf.ca/
[3] http://uqam.ca

2 - Objectifs

Une formation en Web sémantique n'est pas sans défis. La complexité de cette technologie, et les ressources spécialisées que requière son développement en sont les principaux. La complexité de cette technologie nécessite une approche pédagogique particulière, parce qu'elle inclut plusieurs couches technologiques, plusieurs langages, et plusieurs protocoles. Elle nécessite aussi des spécialistes distincts pour les différents composants de cette technologie : essentiellement ce sont des programmeurs, des programmeurs web, des spécialistes en ontologies et représentation des connaissances, des spécialistes en données et des bibliothécaires systèmes.

C'est pour relever ces défis que nous avons, dans le cadre de ce projet, mis en place une approche pédagogique particulière avec un environnement d'apprentissage. L'objectif était double : concevoir un environnement d'apprentissage en Web sémantique, et former des professionnels en Web sémantique. L'environnement d'apprentissage qui est proposé ici est le résultat d'années d'expérience de Michel Héon en recherche et en formation en Web sémantique. Le présent document en explique le fonctionnement, dans le cadre d'une démarche pour bâtir des compétences en Web sémantique. Il en détaille l'approche pédagogique sous-jacente, le plan de formation, le processus et les ressources nécessaires.

La clientèle-cible de cette formation est constituée de bibliothécaire systèmes, de programmeurs d'application et d'interface Web, et de spécialistes en gestion de données. La formation consiste à outiller les participants de notions fondamentales sur le Web sémantique afin qu'ils puissent acquérir les compétences nécessaires au développement d'ontologies au regard des besoins des établissements de recherche.

Ce projet a de multiples impacts anticipés pour la recherche **:** outre pallier à la rareté de compétences en Web sémantique au Canada et bâtir les compétences en IA dont le Canada aura besoin dans un proche avenir, il a pour objet de former du personnel hautement qualifié pour aider à la dissémination et à la valorisation de la recherche canadienne, et, de manière plus générale,

3 - Déroulement du projet

La première phase du projet consistait en cibler la clientèle, déterminer ses besoins en formation, et procéder au recrutement du formateur, puis construire le plan de formation, développer l'outil de formation, et adapter le matériel de formation existants aux besoins du projet. Au cours de cette même phase, nous avons géré les aspects logistiques, qui consistaient en la détermination du calendrier de formation, l'invitation des participants à la formation, la réservation des salles de cours et de laboratoire, et l'invitation des participants. Peu avant la formation, avec l'aide du soutien technique des Services informatiques de l'UQAM, nous avons préparé les salles de formation avec les postes informatiques et le déploiement de l'environnement de développement.

Sur ces bases, la formation que nous avons intitulé « Les fondamentaux du Web sémantique » s'est déroulée la semaine du 9 au 13 mars 2020 en présentiel à l'UQAM, pour 22 professionnels. Voici la distribution des participants selon leurs qualifications : 7 bibliothécaires systèmes, 7 bibliothécaires spécialisés, 3 programmeurs-analystes, 2 analystes en gestion de données, 1 programmeur Web, 1 architecte système, 1 agent de recherche. Cette diversité dans le public-cible de la formation démontre l'interdisciplinarité que requièrent habituellement les développements en Web sémantique. Lors de la formation, nous avons pu constater la diversité des intérêts et des questions posées qui ont permis d'enrichir les explications du formateur à l'ensemble du groupe.

4 - Approche pédagogique

Pour répondre aux défis énoncés plus haut, nous avons privilégié une approche pédagogique où le formateur progresse dans son enseignement en décrivant successivement les couches technologiques du Web sémantique.

Du point de vue de la forme, la formation comporte une partie magistrale, suivis d'ateliers de pratique et d'expérimentation. Les cours magistraux permettent à l'auditeur d'acquérir le cadre théorique nécessaire à l'atteinte des objectifs de la formation. La période de pratique est l'occasion d'appliquer la théorie et de développer les savoir-faire par des mises en situation.

Dans le cadre de cette formation, la plateforme VIVO[4] a servi de support et d'exemple pour l'apprentissage. VIVO est une plateforme en Web sémantique, en code source ouvert, qui permet partager les données ouvertes de la recherche à travers ses ontologies. VIVO vise à gérer un dossier intégré des profils des chercheurs et de leurs activités. Bien sûr, cette formation ne se limite pas à la compréhension de VIVO car elle permet, par extension, de comprendre n'importe quelle application en Web sémantique et les ontologies y afférent. UQAM a implanté une version pilote de la plateforme VIVO en 2019[5].

[4] http://vivoweb.org
[5] http://expertises.uqam.ca

5 - Formation

Nous avons divisé la formation en trois volets correspondants aux objectifs de la formation : un premier volet pour expliquer le Web sémantique et en décrire l'architecture, un deuxième volet pour présenter les ontologies et enseigner les langages qui les sous-tend, et un troisième volet sous la forme d'un atelier pratique pour initier à l'environnement de développement et l'utiliser. Cette structure en trois volets permet de descendre dans des niveaux progressifs de complexité : ainsi, les participants de suivre l'ensemble de la formation, mais aussi de choisir le niveau de complexité selon leur rôle dans le développement.

Volet 1 - Fondamentaux du web sémantique

Dans ce premier volet, l'objectif est de définir et expliquer ce qu'est le Web sémantique et son architecture. Cette explication se fait en trois étapes.

1ère étape : nous commençons par décrire l'architecture du Web, puisque cette technologie est familière aux utilisateurs : nous utilisons tous au quotidien Internet et le Web, cela fait donc partie d'une connaissance partagée et commune. Le Web est une application de l'Internet qui vise à faciliter un échange interopérable des connaissances, des informations et des données. À travers l'histoire du Web, le formateur explique la vision du Web et du Web sémantique.

La technologie du Web sémantique est standardisée par le World Wide Web Consortium (W3C)[6], organisme à but non-lucratif dont la mission est de fixer les standards du Web selon trois principes : gratuité, données non-propriétaires et liberté d'accès. Ces mêmes principes s'appliquent au Web sémantique.

[6] https://www.w3.org

L'architecture du Web est basée sur la notion d'**interopérabilité**, la capacité que possède un système informatique à échanger des données avec d'autres produits ou systèmes informatiques, existants ou futurs, indépendante de la technologie des systèmes d'informations. L'information doit se diffuser indépendamment de la plateforme technologique, quel que soit l'ordinateur, le système d'exploitation, ou le navigateur Web.

Il faut rappeler les trois composants sur lesquels la technologie du Web repose : (1) un standard d'adressage des ressources (figure 1) : International/Uniform resource Identifier (IRI/URI) (2) un protocole de communication : HyperText Transfer Protocol (HTTP) et (3) un langage de représentation pour les documents : HyperText Modeling Language (HTML). Cette technologie est opérationnalisée selon le modèle Client-Serveur.

2ème étape : le formateur se base sur la description précédente pour expliquer l'architecture du Web sémantique (ou Web 3.0), L'architecture du Web sémantique repose sur la technologie du Web : le Web classique met en lien des documents, alors que le Web sémantique part des mêmes principes pour mettre en lien des données. Le Web permet de transporter des documents de manière interopérable (via des pages Web), le Web sémantique permet de transporter des données de manière interopérable.

Le Web sémantique reprend les trois composants du Web énoncés précédemment en étendant le 3ème ainsi : (3) le langage de représentation pour les documents du Web classique (HTML) est enrichi d'un langage de représentation pour les données : le Resource Description Framework (RDF)

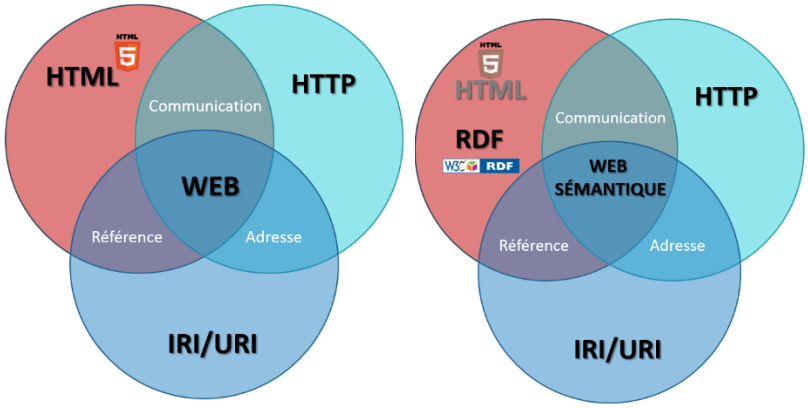

Figure 1: architecture du Web vs. architecture du Web sémantique

Ensuite, nous présentons l'architecture technologique du web de données ouvertes et liées : avec l'importance que prennent les données ouvertes dans le monde de la recherche, la technologie du Web sémantique se révèle un potentiel énorme. En plus de partager les données, le Web sémantique permet de transmettre leur sens (sémantique) ce qui permet une interopérabilité par une machine, ce qui en fait de l'Intelligence Artificielle.

Nous faisons suivre cette explication par la présentation des différents cas d'usage du Web sémantique.

Comment se fait la communication dans le web de données ? L'étape suivante consiste à expliquer la pile langagière du Web sémantique. Dans la figure 2, nous voyons à gauche la pile des différents langages, et à droite quelles fonctionnalités ont ces langages.

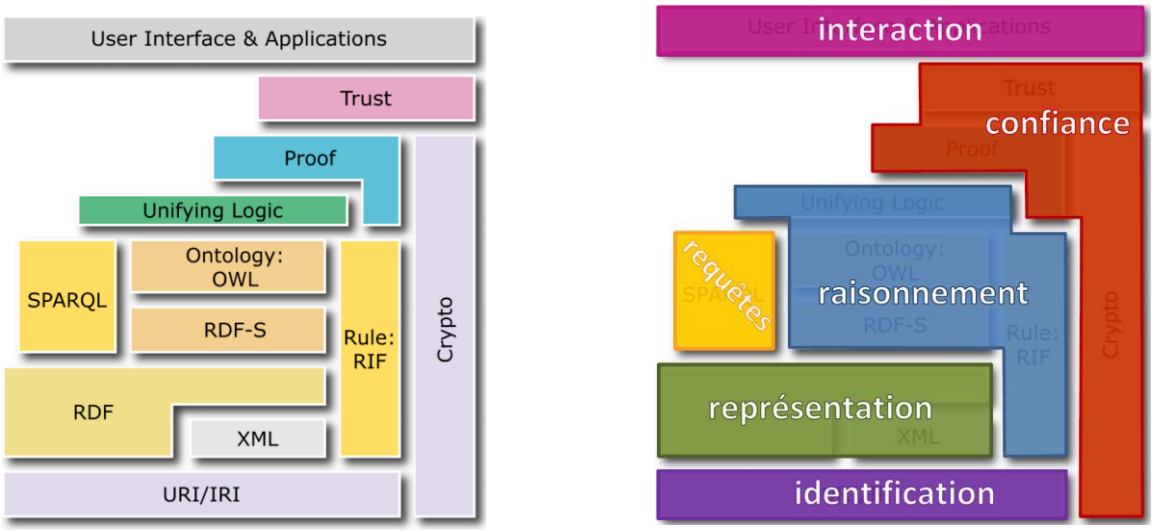

Figure 2: la pile langagière du Web sémantique : langages et fonctionnalités

La description de ces langages est illustrée par l'historique de leur mise en place progressive. La figure 3 donne une bonne idée de l'évolution du Web sémantique depuis 2005.

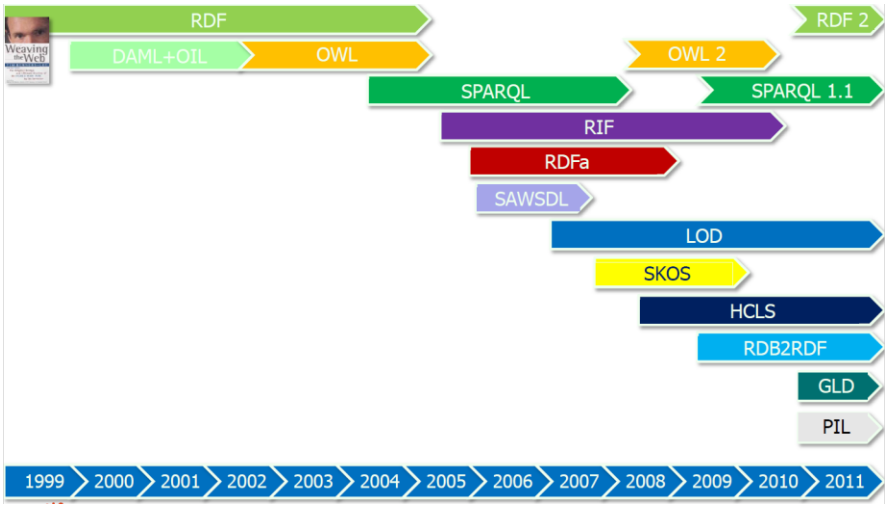

Figure 3: évolution des langages du Web sémantique

Avec ces bases, la formation peut se poursuivre par la description complète du modèle de données en Web sémantique, qui est résumé sous une forme graphique à la figure 4.

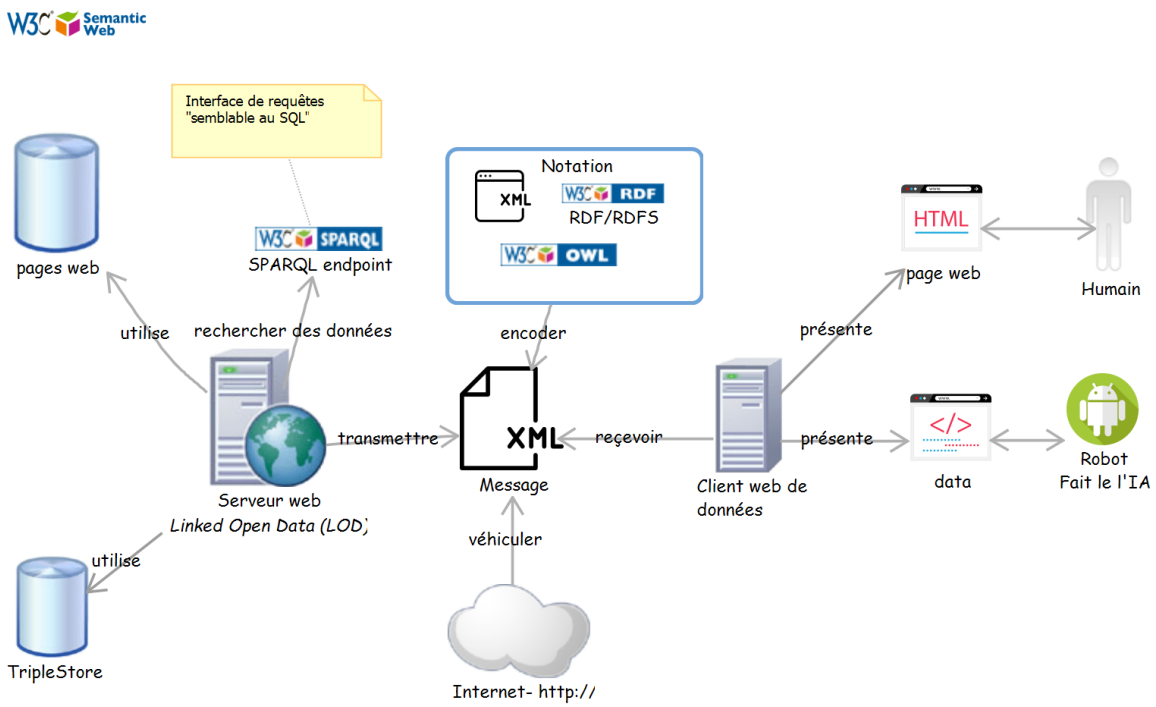

Figure 4: modèle de données en Web sémantique

3ème étape : comme application de ce modèle et pour montrer un exemple, le formateur explique l'architecture de VIVO. La figure 5 présente l'architecture Trois tiers (en trois parties). À sa base se trouve le regroupement de technologies servant à la pérennisation des données (SOLR, MySQL ou le TripleStore Jena). La 2ème couche comporte les technologies utiles à l'implantation de la logique métier. La 3ème couche présente les technologies utilisées pour présenter les données à l'utilisateur. Le participant aura ainsi acquis les notions nécessaires à la compréhension du fonctionnement global de VIVO, et des applications en Web sémantique en général.

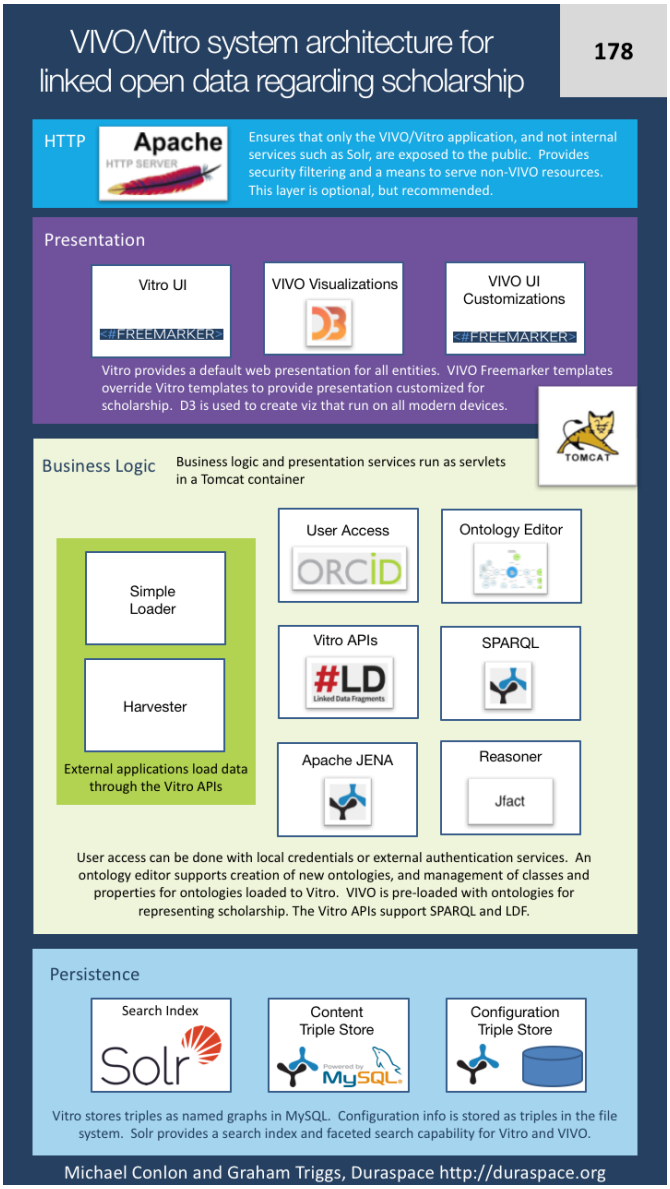

Figure 5: architecture de VIVO

Volet 2 – Modélisation d'ontologies

Le deuxième volet porte sur la modélisation des ontologies. Cette partie est sans doute la plus complexe de la formation, mais probablement aussi la plus intéressante. C'est la partie qui intéresse en particulier les libraires systèmes (rappelons que le public-cible de cette formation est en partie constitué de libraires systèmes, étant donné le potentiel que les données ouvertes liées présentent pour la gestion des documents et des données).

Dans ce volet sont abordées les notions d'ontologies et de représentation des connaissances, de structure des connaissances, des modèles de connaissance.

Sur ces bases, nous décrivons les langages des ontologies, notamment le Web Ontology Language (OWL). Nous expliquons les caractéristiques d'une notation, comment nous structurons une conceptualisation dans l'ontologie, et comment nous modélisons une ontologie. La terminologie qui sert de base aux ontologies, et la structure de données du Web sémantique (sémantique, vocabulaires, taxonomies) est définie, expliquée, et illustrée par des exemples. Le web de données ouvertes et liées (Linked Open Data - LOD). Nous montrons un exemple avec DBPedia, l'application phare du Web sémantique et du LOD qui structure les données source pour Wikipédia.

Ensuite, nous expliquons la notion de ressources, et comment elles sont identifiées dans le Web. Le schéma de description des ressources : le *Resource Description Framework (RDF),* et son Schéma : *RDF-Schema (RDFS).*

À la suite de cela, nous décrivons les requêtes SPARQL qui permettent de rechercher, d'extraire des données des ontologies, et comment sont structurées ces requêtes. Quelques exemples sont montrés. Lors du troisième volet de la formation, il y aura plusieurs exercices avec des requêtes SPARQL dans l'environnement d'apprentissage.

VIVO est utilisé comme exemple pour montrer les ontologies dans une application (figure 6). VIVO a réuni un certain nombre d'ontologies existantes dans une structure sémantique unifiée. Il y a des ontologies génériques : Friend of a Friend (FOAF) qui sert à lier les personnes entre eux, et qui est utilisées notamment dans le réseau social Facebook, Event Ontology, SKOS (Simple Knowledge Organization System) et vCard. Plusieurs autres ontologies dans VIVO ont été créées par OBO

l'Open Biological and Biomedical Ontology (OBO) Foundry[7], dont la mission est de créer des ontologies interopérables pour la science. Enfin, il est possible de développer des extensions aux ontologies existantes, ou d'en ajouter de nouvelles.

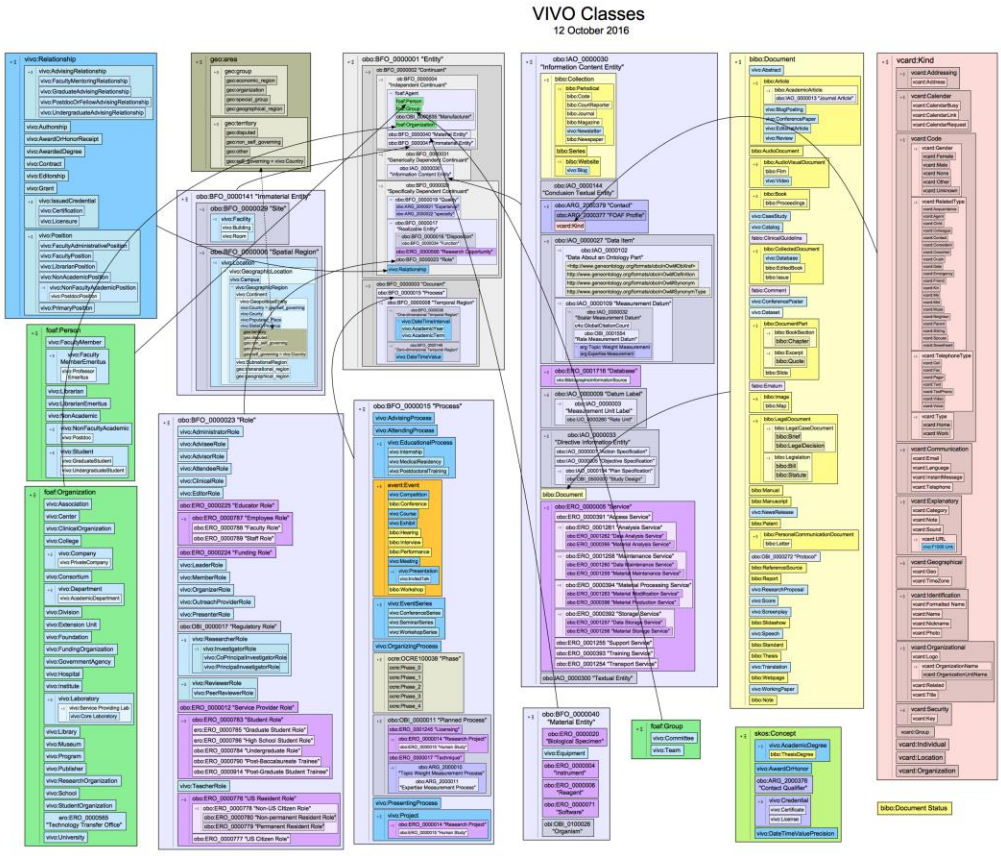

Figure 6: classes représentant les ontologies de VIVO

Volet 3 - Environnement d'apprentissage[8]

Le troisième volet consiste en un atelier pratique pour s'approprier l'environnement de développement conçu dans le cadre de ce projet. Cet environnement, que nous avons nommé UQAM-DEV, est l'aboutissement du développement d'un environnement complet pour le développement en web sémantique.

[7] http://obofoundry.org

[8] Cet environnement de développement a été présenté lors de la conférence internationale VIVO en juin 2020. L'outil est accessible via les liens présents sur le Wiki public de l'UQAM : https://wiki.uqam.ca/

L'avantage de cet environnement est qu'il est adapté à une utilisation par des non-programmeurs, (un des défis relevés au début du projet), car le Web sémantique requiert des spécialistes ayant des bagages différents (notamment des bibliothécaires, et des analystes en gestion de données). Bien qu'il soit aussi utilisé par des programmeurs, l'environnement doit donc rester accessible aux non-programmeurs (comme les spécialistes en ontologie, les bibliothécaires système, ou les gestionnaires de projet).

UQAM-DEV est un environnement de développement complet, qui permet de de compiler, installer, déployer une instance d'une application en Web sémantique sur un ordinateur. En outre, il permet de modéliser, d'éditer les ontologies dans les notations existantes reconnues par le W3C.

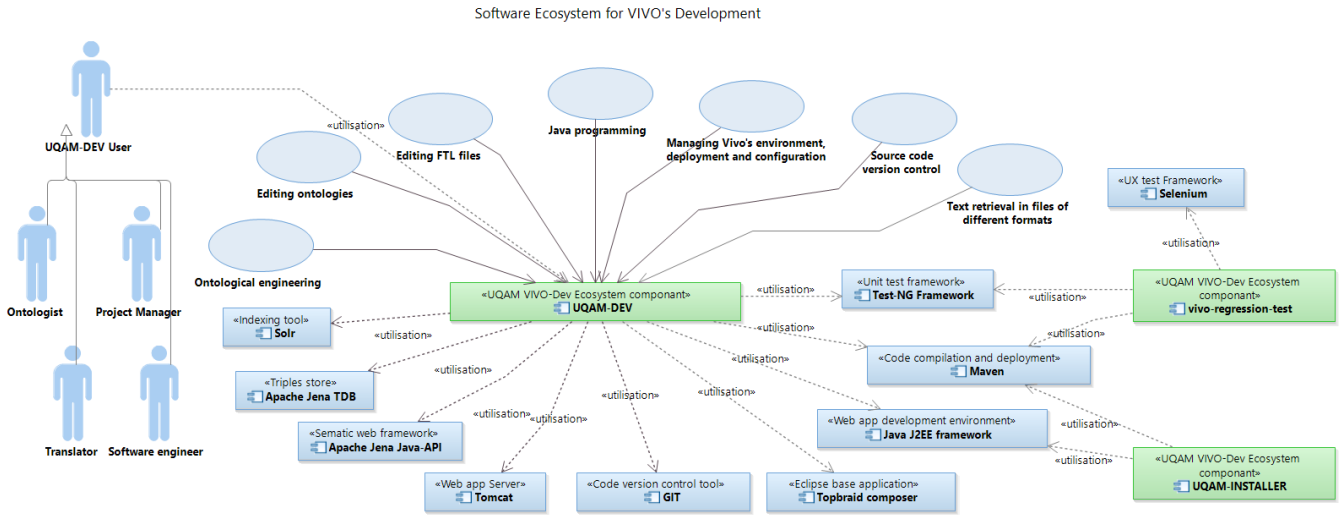

Figure 7 : composants de UQAM-DEV

L'utilité de cette environnement intégré de développement est de favoriser le travail d'une équipe composée de programmeur, de spécialistes en ontologies, de gestionnaires de projets. Il constitue un écosystème évolutionnaire et adaptif, car d'autres outils peuvent y être intégrés au fur et à mesure des besoins, et le flux de travail peut en être amélioré. Il permet le développement d'applications en Web sémantique ou l'édition d'ontologies, en alignement avec la méthodologie de développement Agile.

Du point de vue technique, l'environnement UQAM-DEV est construit sur la base de Topbraid Composer[9], un éditeur d'ontologie conçu pour être intégré à Eclipse[10], l' environnement de développement Java le plus utilisé. UQAM-DEV permet d'éditer les ontologies dans les notations existantes reconnues par le W3C (.n3; .ttl; RDF/XML), et il intègre les fonctionnalités de compilation Java, Maven[11], J2EE, Tomcat[12], et le contrôle de version (figure 7). L'environnement intègre une instance de Tomcat qui contient l'application, une instance de SOLR qui gère l'index de recherche, et un ensemble de macros pour réaliser les tâches requises pour le développement en Web sémantique et l'édition d'ontologies.

La présentation de l'environnement UQAM-DEV se fait dans une approche progressive. Après un tour général de ses fonctionnalités, nous procédons à une description de l'espace de travail. Nous montrons comment l'environnement permet de travailler avec du code sur un espace commun Git[13], comment fonctionne l'extraction, l'ajout, ou la modification de code, et la notion de branche dans Git, comment éditer le code, et comment éditer les ontologies. Nous indiquons aussi comment se fait la navigation dans les fichiers gérés dans l'environnement de développement : dans les répertoires de code source, dans Tomcat, comment se fait la gestion des fichiers de propriétés et des fichiers d'ontologies (figure 8).

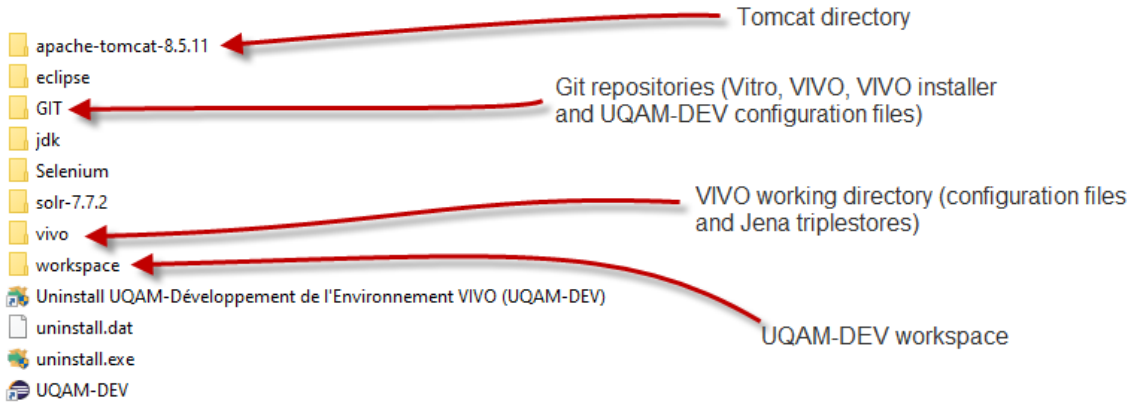

Figure 8: structure de répertoires de UQAM-DEV

[9] https://www.topquadrant.com/products/topbraid-composer/
[10] https://www.eclipse.org/ide
[11] https://maven.apache.org
[12] http://tomcat.apache.org
[13] https://git-scm.com

Comme exemple d'utilisation, nous présentons le cycle de développement standard d'une application en Web sémantique (figure 8), avec en exemple l'application VIVO. Nous montrons comment ce processus de développement est opérationnalisé dans l'environnement UQAM-DEV.

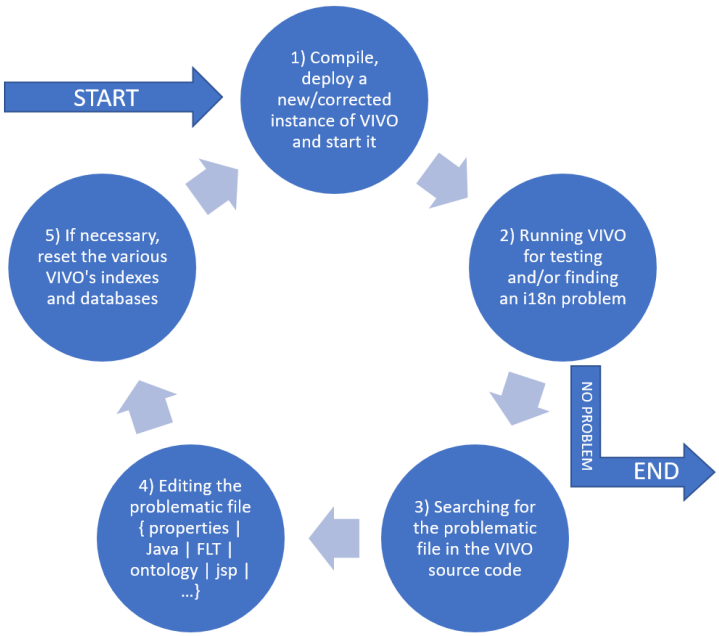

Figure 8: cycle de développement d'une application en Web sémantique (exemple de VIVO)

Présentée à la figure 9, l'interface se présente sous la forme d'un ensemble de ressources visuelles dans quatre catégories : des perspectives (1), des menus (2), les barres d'outil (3) et les vues (4) que nous détaillons ci-dessous :

1. **Les perspectives** permettent à l'utilisateur d'utiliser les fonctionnalités de l'environnement de développement selon la perspective que requiert de son rôle au moment du développement. Dans UQAM-DEV, cinq perspectives sont offertes : celle du spécialiste en ontologies (Topbraid) celle du programmeur (Java), celle du testeur (Debug), ainsi que la perspective pour l'utilisation du dépôt (GIT) et celle des ressources. Cinq perspectives ont donc disponibles.
2. **Les menus** correspondent à des ensembles d'actions qui peuvent être entreprises à l'intérieur de ces perspectives : les menus changent en fonction de la perspective choisie.
3. **Les barres d'outils** : adaptable aux perspectives, elles ont des icônes pour les fonctions les plus utilisées communément.
4. **Les vues :** fenêtres qui représentent de parties de l'information selon la perspective choisie.

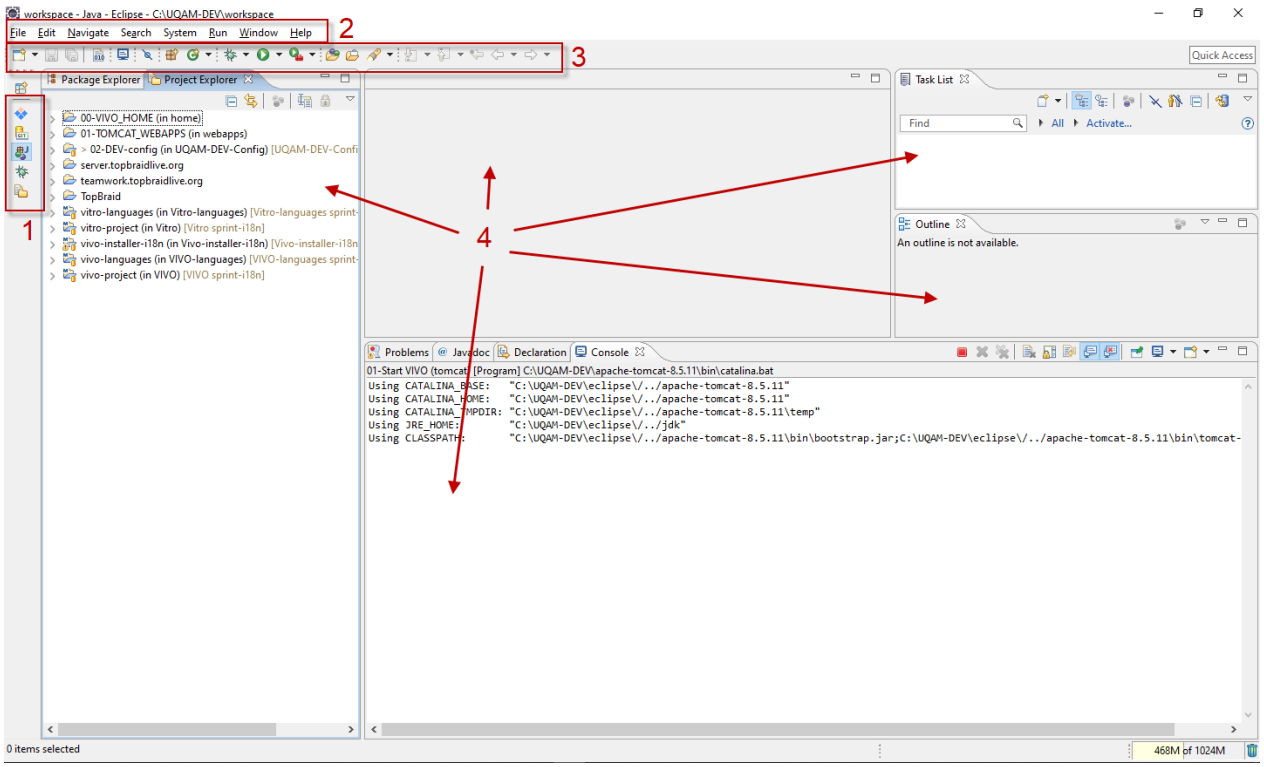

Figure 9: interface de UQAM-DEV

À partir de ces explications sur les fonctionnalités et les technologies d'UQAM-DEV, de ce contexte (le développement de VIVO), et de la présentation de l'interface, nous pouvons montrer comment acquérir le code de l'application, compiler celle-ci, l'installer dans Tomcat, et la tester. Toutes ces étapes se font à travers plusieurs exercices dans l'atelier, afin que les participants l'expérimentent de première main. C'est la partie la plus appliquée de la formation.

6 - Résultats, limites et futurs développements

La formation réalisée avec l'approche pédagogique par « couches successives » et l'environnement d'apprentissage que nous avons décrit, ont portées leurs fruits. Après quelques jours de formations, les assistants ont eu une positive d'avoir appris beaucoup, tout en réalisant la complexité de la technologie. Les participants ont été très satisfaits de la formation qui leur a permis d'acquérir des compétences rares dans une période relativement courte, et la complexité de la technologie présentée a été présenté d'une manière qui favorisait son apprentissage.

Bien sûr, quelques jours de formation ne suffisent pas à maîtriser une technologie aussi complexe, et ne remplacent pas une expérience de quelques années. C'est pourquoi une partie des personnes formés travaillent sur le projet VIVO et ont la possibilité de mettre leur apprentissage à contribution et de ne pas perdre les bénéfices de ces quelques jours de formation intensive.

Pour les participants, il y a cependant un besoin d'approfondir les connaissances. Nous souhaitons développer des modules plus spécialisés, selon les publics-cibles et les compétences nécessaires, dans les mois prochains, afin de ne pas perdre les acquis de cette première formation.

Bibliographie

Dimitrov, M. (2012). Semantic Technologies and Triplestores for Business Intelligence. In M.-A. Aufaure & E. Zimányi (Eds.), Business Intelligence: First European Summer School, eBISS 2011, Paris, France, July 3-8, 2011, Tutorial Lectures (pp. 139-155). Berlin, Heidelberg: Springer Berlin Heidelberg.

Sikos, L. F. (2015). Knowledge Representation *Mastering Structured Data on the Semantic Web: From HTML5 Microdata to Linked Open Data* (pp. 13-57). Berkeley, CA: Apress.

Sikos, L. F. (2015). Linked Open Data *Mastering Structured Data on the Semantic Web: From HTML5 Microdata to Linked Open Data* (pp. 59-77). Berkeley, CA: Apress.

Yu, L. (2014). FOAF: Friend of a Friend *A Developer's Guide to the Semantic Web* (pp. 357-382). Berlin, Heidelberg: Springer Berlin Heidelberg.

Yu, L. (2014). Other Recent Applications: data.gov and Wikidata *A Developer's Guide to the Semantic Web* (pp. 551-585). Berlin, Heidelberg: Springer Berlin Heidelberg.