Université du Québec à Montréal

Research and Creation Department

THE SEMANTIC WEB AND OPEN DATA AS LEVERS OF IMPACT IN RESEARCH: FUTURE SKILLS IN ARTIFICIAL INTELLIGENCE

Michel Héon PhD, Development and Training

Rachid Belkouch, Project Management

1st semester 2020

Future Skills Centre | Centre des Compétences futures

RESEARCH IMPACT CANADA | Turning research into action

Table of Contents

List of Illustrations

1 - Introduction

This project focuses on the development of a Semantic Web-based training tool for research data management personnel consisting of a learning environment and a training guide.

The emergence of open data in research is encouraging universities to improve the visibility, promotion and dissemination of research - in other words, the impact of research. Semantic Web technologies make it possible to link open data across ontologies. This possibility is especially important in today's world of research, where cooperation has become so significant, be it between researchers themselves, or between research institutions and industry.

The Semantic Web is also a proven AI (Artificial Intelligence) technology. AI is unquestionably a professional skill that will be required in the future, as there is a significant lack of human resources in Canada dedicated to the Semantic Web.

Thanks to funding from Research Impact Canada (RIC)[1] and the Future Skills Centre (FSC)[2], the Université du Québec à Montréal (UQAM)[3] developed a learning environment and conducted training for a cohort of 22 professionals during the first semester of 2020.

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

2 - Goals

Semantic Web training is not without its challenges. The complexity of this technology and the specialized resources required for its development are the main obstacles. The complexity of this technology requires a particular educational approach because it includes several technological layers, several languages, and several protocols. It also requires a distinct set of specialists for the various components of this technology: essentially programmers, web programmers, as well as ontology and knowledge representation specialists, data specialists, and systems librarians.

In order to overcome these challenges, we have, within the scope of this project, implemented a particular educational approach with a learning environment. The objective was twofold: to design a Semantic Web learning environment, and to train professionals in the Semantic Web. The learning environment presented in this document stems from Michel Héon's years of experience in research and training in the Semantic Web. This document outlines its functioning, as part of an approach to developing skills related to the Semantic Web. It further specifies the underlying educational approach, the training plan, the process and the resources required.

The target audience for this training consists of systems librarians, application and web interface programmers, as well as data management specialists. The training consists of providing participants with the fundamental notions of the Semantic Web so that they can acquire the skills needed to develop ontologies that meet the needs of research institutions.

This project has a number of anticipated impacts for research**:** in addition to addressing the scarcity of Semantic Web skills in Canada and building the AI skills that Canada will need in the near future, it is intended to train highly qualified personnel to assist in disseminating and promoting Canadian research, in a broader sense.

3 - Progress of the Project

The first phase of the project consisted of targeting the clientele, determining its training needs, and recruiting the trainer, followed by creating the training plan, developing the training tool, and adjusting the existing training materials to meet the needs of this project. During this phase, we managed the logistics, which consisted of determining the training schedule, inviting participants to the training, booking the classrooms and labs, as well as issuing invitations to the participants. Shortly before the training, we prepared the training facilities with the computer workstations and deployed the development environment, with the help of the technical support team of UQAM's IT Services.

Accordingly, the training we named "The Fundamentals of the Semantic Web" was held during the week of March 9-13, 2020, in a face-to-face course at UQAM involving 22 professionals. The distribution of participants according to their qualifications was as follows: 7 system librarians, 7 specialized librarians, 3 programmer-analysts, 2 data management analysts, 1 web programmer, 1 system architect, 1 research officer. Such diversity within the training's target audience reflects the interdisciplinary nature usually required for Semantic Web developments. Throughout the training, we noted a wide range of interests and a diversity of questions that enriched the trainer's explanations to the entire group.

4 - Educational Approach

To meet the challenges stated above, we preferred an educational approach in which the trainer moves through his teaching by sequentially describing the technological layers of the Semantic Web.

In terms of approach, the training consists of a lecture part, followed by practical and experimental workshops. The lectures enable the listener to acquire the theoretical framework needed to attain the goals of the training. The practice period is an opportunity to apply the theory and develop know-how through simulations.

The VIVO platform[4] was used as an example and medium for learning during this training. VIVO is an open source Semantic Web platform that allows sharing open research data across its ontologies. VIVO aims to manage an integrated record of researchers' profiles and their activities. Of course, this training is not limited to understanding VIVO because, by extension, it allows you to understand any Semantic Web application and its related ontologies. UQAM implemented a pilot version of the VIVO platform in 2019[5].

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

5 - Training

We have divided the training into three components corresponding to the training objectives: the first component explains the Semantic Web and describes its architecture, the second component presents the ontologies and teaches the underlying languages, and the third component is a practical workshop to introduce the development environment and use it. This three-part structure allows participants to progressively delve into levels of complexity, enabling them to follow the entire training, but also to choose the level of complexity according to their role in development.

Part 1 - The Fundamentals of the Semantic Web

In this first section, the focus is to define and explain the Semantic Web and its architecture. This explanation is done in three steps.

1st step: we begin by describing web architecture, since this technology is familiar to users: we all use the Internet and the web on a daily basis, so it is part of a shared and common knowledge. The web is an Internet application whose purpose is to facilitate an interoperable exchange of knowledge, information and data. Using the history of the web, the trainer explains the concept of the web and the Semantic Web.

The Semantic Web technology is standardized by the World Wide Web Consortium (W3C)[6], a non-profit organization whose mission is to set web standards in accordance with three principles: free, non-proprietary data and freedom of access. These same principles apply to the Semantic Web.

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

Web architecture is based on the concept of **interoperability**, the ability of a computer system to exchange data with other products or computer systems, existing or future, independent of information systems technology. Information must be disseminated regardless of the technological platform, computer, operating system, or web browser.

It is worth remembering the three components on which web technology is based: (1) a resource addressing standard (Figure 1): International/Uniform Resource Identifier (IRI/URI); (2) a communication protocol: HyperText Transfer Protocol (HTTP); and (3) a document representation language: HyperText Modeling Language (HTML). This technology is operationalized according to the Client-Server model.

2nd step: the trainer uses the previous description to explain the architecture of the Semantic Web (or Web 3.0). The architecture of the Semantic Web is based on web technology: the traditional web connects documents, while the Semantic Web connects data using the same principles. The web enables documents to be transmitted in an interoperable way (via web pages), while the Semantic Web enables data to be transmitted in an interoperable manner.

The Semantic Web includes the three previously mentioned features of the web and expands the third one thus: (3) the representation language of documents in the traditional Web (HTML) is enhanced with a representation language for data: the Resource Description Framework (RDF).

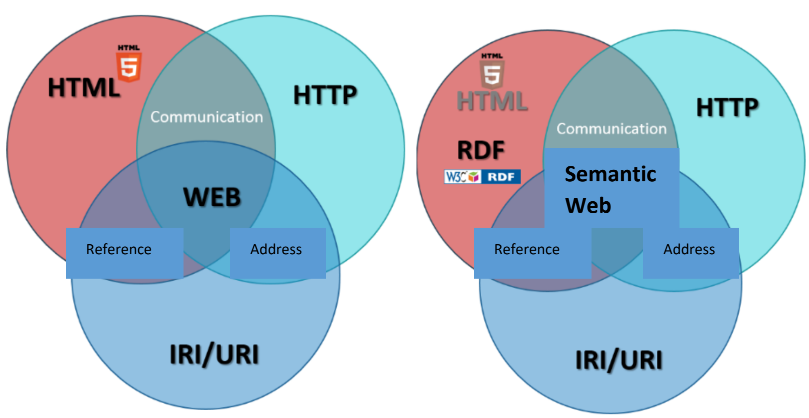

Figure 1: Web Architecture vs. Semantic Web Architecture

	Reference
	Communication
	Address
	SEMANTIC WEB
	IRI / URI

Then, we present the technological architecture of the open and linked data web: with the prominence of open data in the world of research, the Semantic Web technology is proving to have enormous potential. In addition to sharing data, the Semantic Web makes it possible to transmit their (semantic) meaning, which allows interoperability by a machine, making it Artificial Intelligence.

This explanation is followed by a presentation of the different uses of the Semantic Web.

How is data communication done on the web? The next step is to explain the linguistic layers of the Semantic Web. In Figure 2, the left side shows the various language layers, and the right side shows which functionalities these languages have.

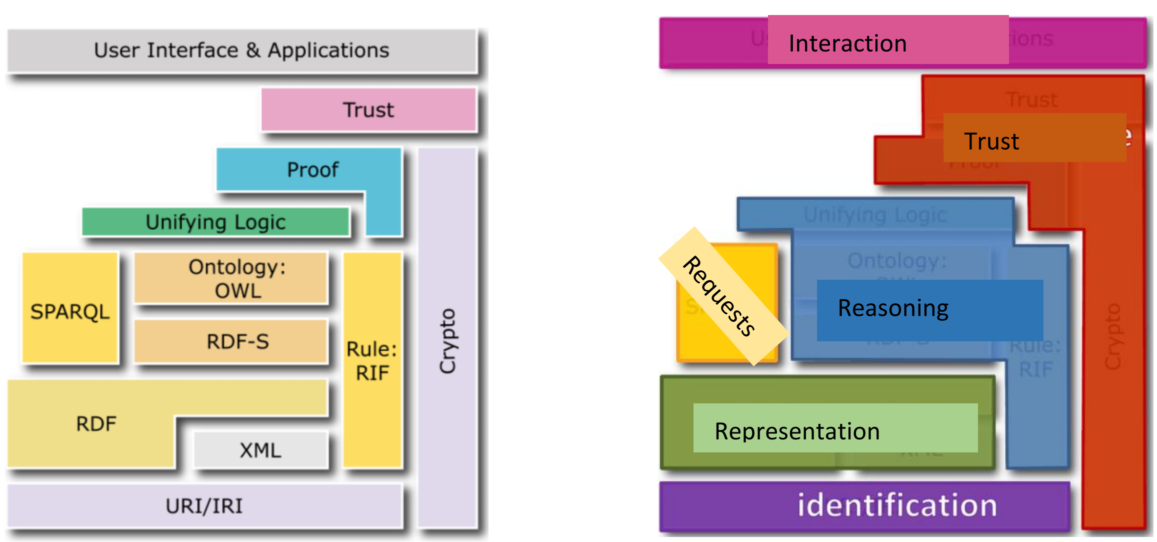

Figure 2: The Linguistic Layers of the Semantic Web: Languages and Functionalities

	Interaction

	Trust
	Reasoning
	Representation
	Identification
	Requests

The description of these languages is presented with a timeline of their gradual implementation. Figure 3 provides a clear picture of the development of the Semantic Web since 2005.

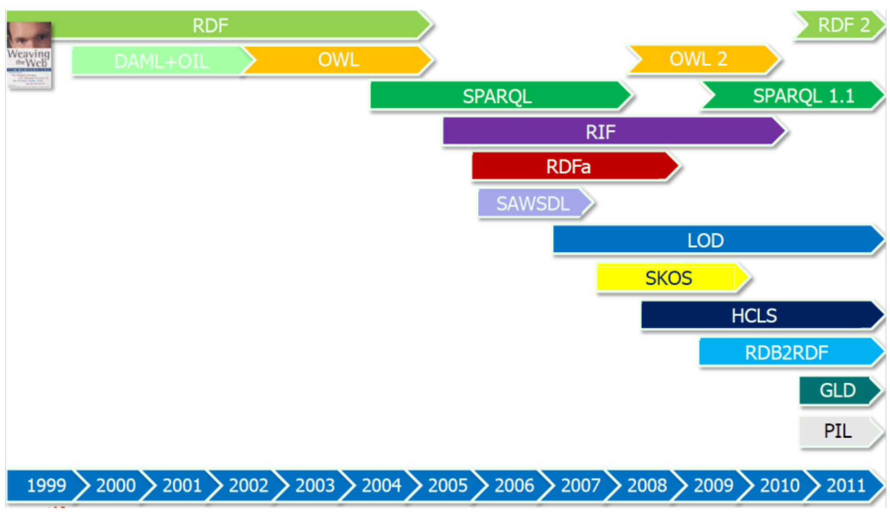

Figure 3: Development of the Semantic Web Languages

Once these fundamentals have been discussed, the training continues with a full description of the Semantic Web data model, which is summarized in the graph in Figure 4.

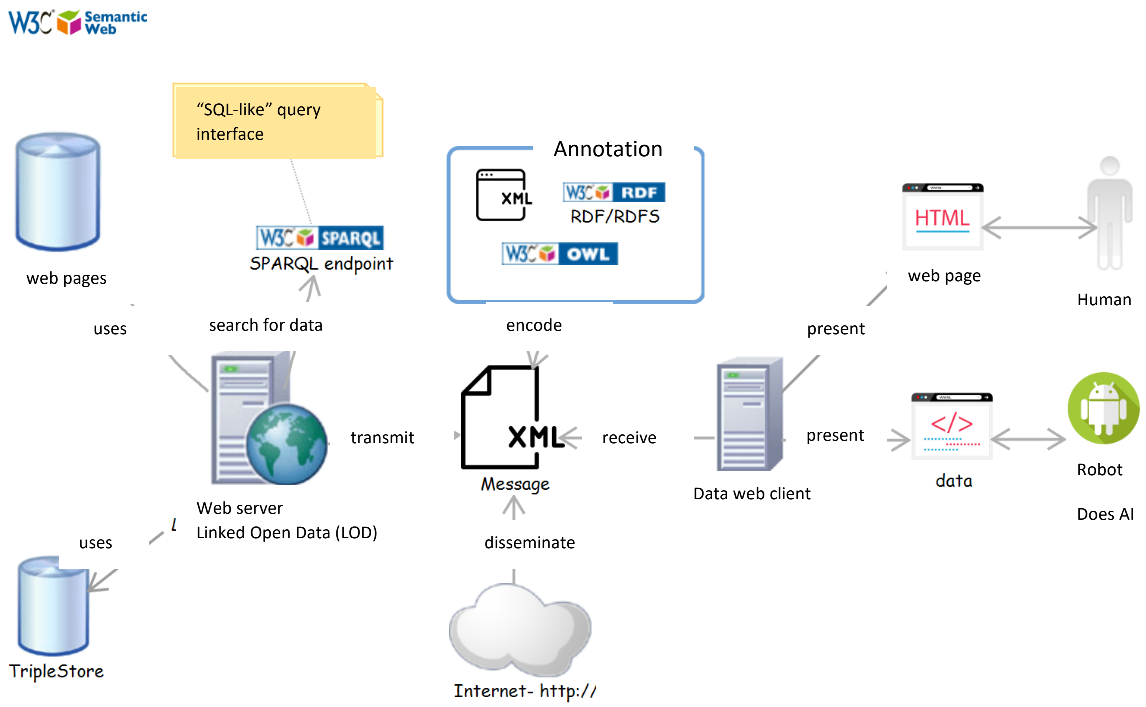

Figure 4: Semantic Web Data Model

	Web pages
	Uses
	Search for data
	Web server
	"SQL-like" query interface
	Transmit
	Annotation
	encode
	Message
	Disseminate
	Receive
	Data web client
	Present
	Data
	Web page
	Robot
	Does AI

3rd step: as an application of this model and to provide an example, the trainer explains the VIVO architecture. Figure 5 shows the three-part architecture. At its base lies the grouping of technologies used to perpetuate data (SOLR, MySQL or the TripleStore Jena). The 2nd layer contains the technologies used to implement the business logic. The 3rd layer features the technologies used to present the data to the user. The participant will thus have acquired the concepts needed to understand the overall functioning of VIVO, as well as Semantic Web applications in general.

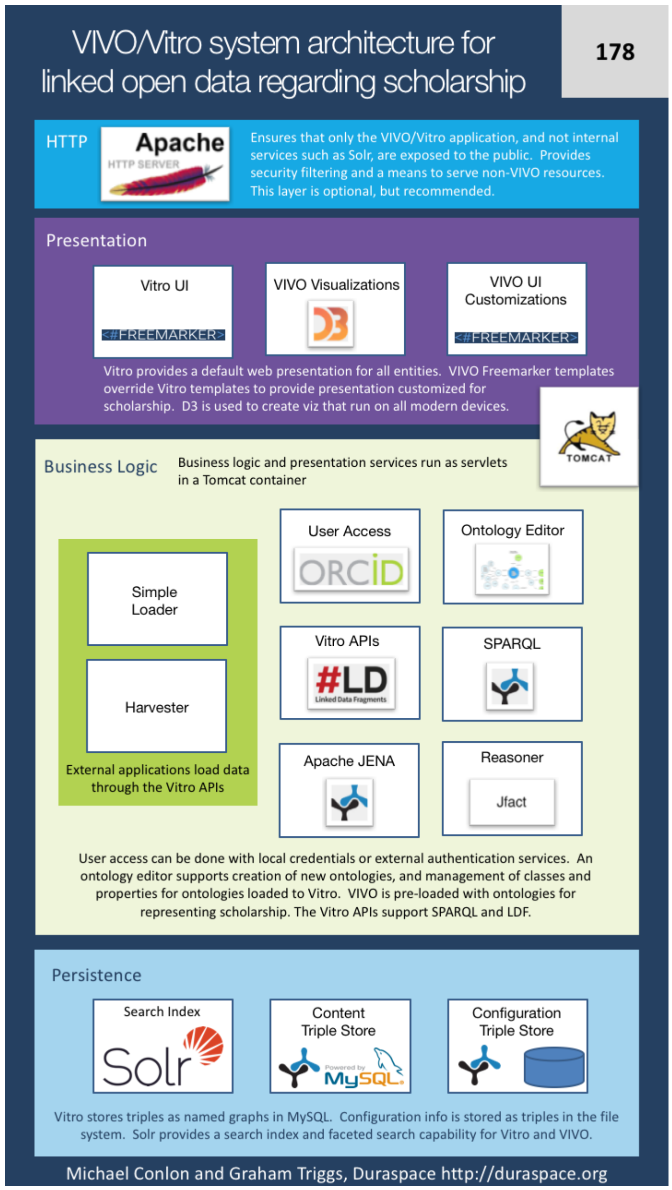

Figure 5: VIVO Architecture

Part 2 - Ontology modeling

The second component deals with ontology modeling. This part is unquestionably the most complex part of the training, but probably also the most interesting one. This section is of particular interest to system librarians (remember that this training's target audience is partly made up of system librarians, given the potential of linked open data for document and data management).

In this section, the concepts of ontologies and knowledge representation, knowledge structures, knowledge models are discussed.

On these bases, we describe the languages of ontologies, particularly the Web Ontology Language (OWL). We explore the characteristics of an annotation, how a conceptualization is structured in ontology, and how to model an ontology. The terminology underlying ontologies, and the data structure of the Semantic Web (semantics, vocabularies, taxonomies) are defined, explained, and illustrated using examples. The Linked Open Data (LOD) web. An example is given using DBPedia, the flagship application of the Semantic Web and LOD that structures source data for Wikipedia.

Next, we explain the concept of resources, and how they are identified on the web. The *Resource Description Framework (RDF),* and its Schema: RDF-Schema *(RDFS).*

Following that, we describe the SPARQL queries that make it possible to search for and extract data from ontologies as well as the structure of these queries. A few examples are given. The third part of the training will include several exercises using SPARQL queries in the learning environment.

VIVO is used as an example to demonstrate ontologies in an application (Figure 6). VIVO has brought together a number of existing ontologies into a unified semantic structure. There are generic ontologies: Friend of a Friend (FOAF), which is used to link people together, applied in particular in the social network Facebook, Event Ontology, SKOS (Simple Knowledge Organization System) and vCard. Several other ontologies in VIVO have been created by OBO, the Open Biological and Biomedical Ontology Foundry[7], whose mission it is to create interoperable ontologies in the field of science. And, finally, it is possible to develop extensions to existing ontologies, or to add new ones.

[7] http://obofoundry.org

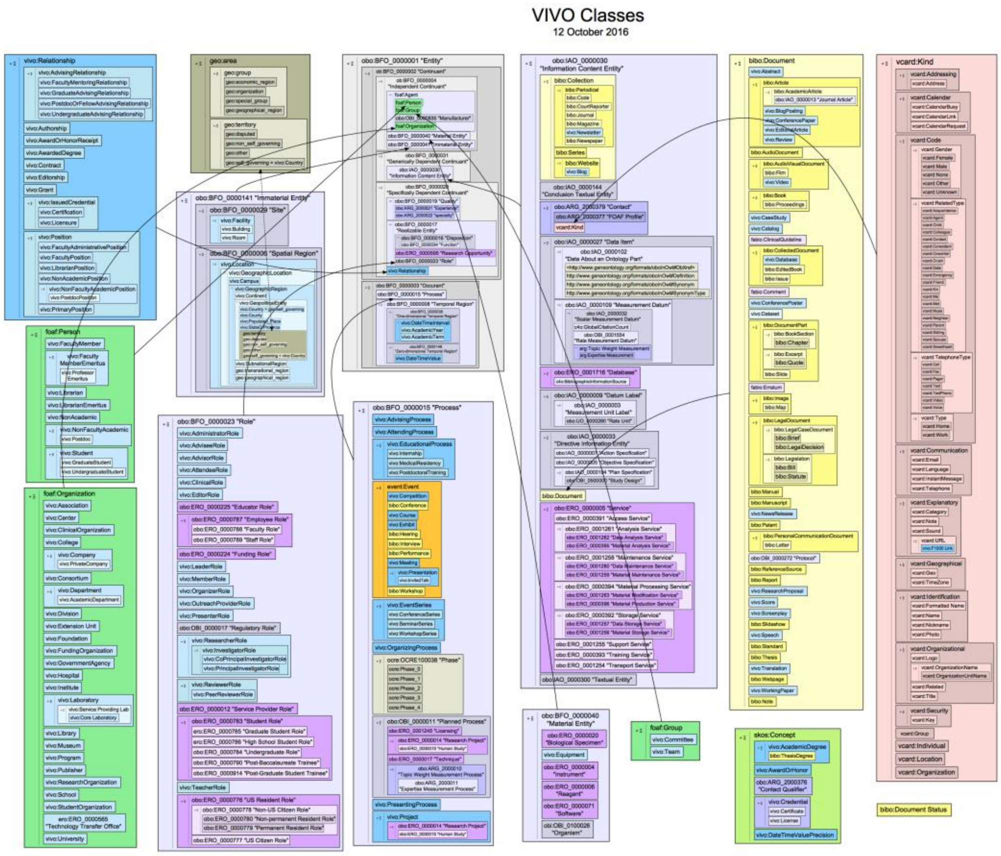

Figure 6: Classes Representing VIVO Ontologies

Part 3 - The Learning Environment[8]

The third component consists of a practical workshop to adapt to the development environment designed within the framework of this project. This environment, which we named UQAM-DEV, is the culmination of the development of a complete environment for Semantic Web development.

The advantage of this environment is that it is suitable for use by non-programmers (one of the challenges raised at the start of the project), because the Semantic Web requires specialists with different backgrounds (including librarians and data management analysts). Although it is also used by programmers, the environment must therefore remain accessible to non-programmers (such as ontology specialists, system librarians, or project managers).

[8] This development environment was presented at the international VIVO conference in June 2020. The tool is accessible via the links on the UQAM Public Wiki: https://wiki.uqam.ca/

UQAM-DEV is a complete development environment, which allows compiling, installing, and deploying an instance of a Semantic Web application on a computer. In addition, it makes it possible to model and edit ontologies in existing annotations recognized by the W3C.

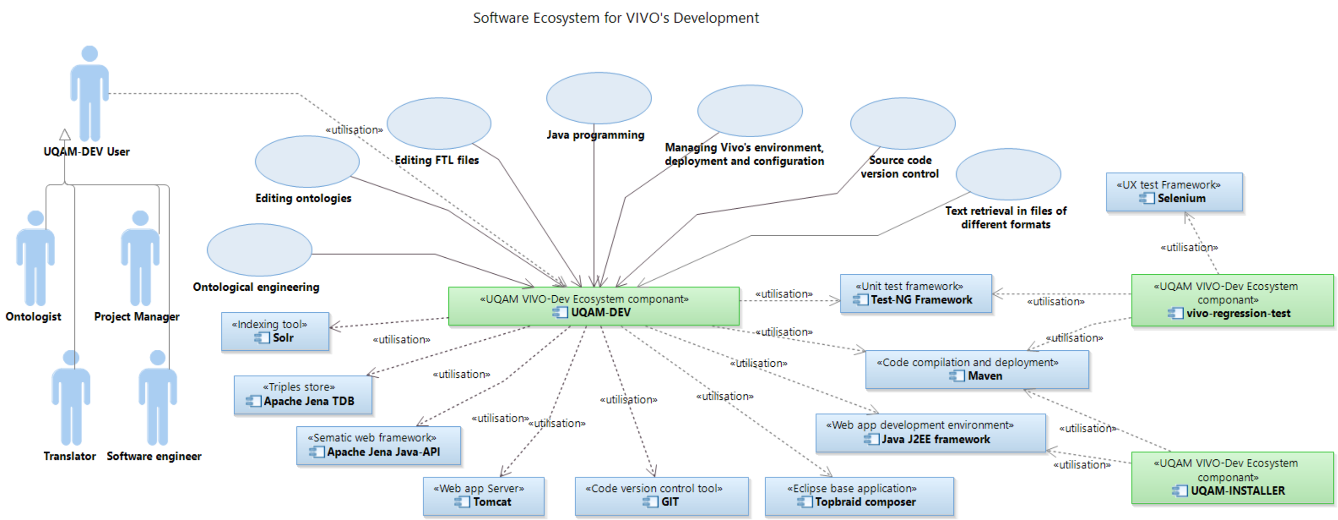

Figure 7 : UQAM-DEV Components

The purpose of this integrated development environment is to foster the work of a team made up of programmers, ontology specialists and project managers. It constitutes an evolutionary and adaptive ecosystem, since other tools can be integrated into it as needed, and the workflow can be improved. It allows the development of Semantic Web applications or the editing of ontologies, in line with the Agile development methodology.

From a technical point of view, the UQAM-DEV environment is built on Topbraid Composer[9], an ontology editor designed to be integrated with Eclipse[10], the most widely used Java development environment. UQAM-DEV makes it possible to edit ontologies in existing annotations recognized by W3C (.n3; .ttl; RDF/XML), and it integrates the Java, Maven[11], J2EE, Tomcat[12] compilation functionalities and version control (Figure 7). The environment integrates an instance of Tomcat

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

that contains the application, an instance of SOLR that manages the search index, and a set of macros to perform the tasks required for Semantic Web development and ontology editing.

A progressive approach is used to introduce the UQAM-DEV environment. After a general tour of its functionalities, we proceed with a description of the workspace. We demonstrate how the environment allows users to work with code on a common Git workspace[13], how to extract, add, or modify code, and the concept of branch in Git, how to edit code, and how to edit ontologies. We also show how the files managed in the development environment are browsed: in the source code directories, in Tomcat, how the property files and ontology files are managed (Figure 8).

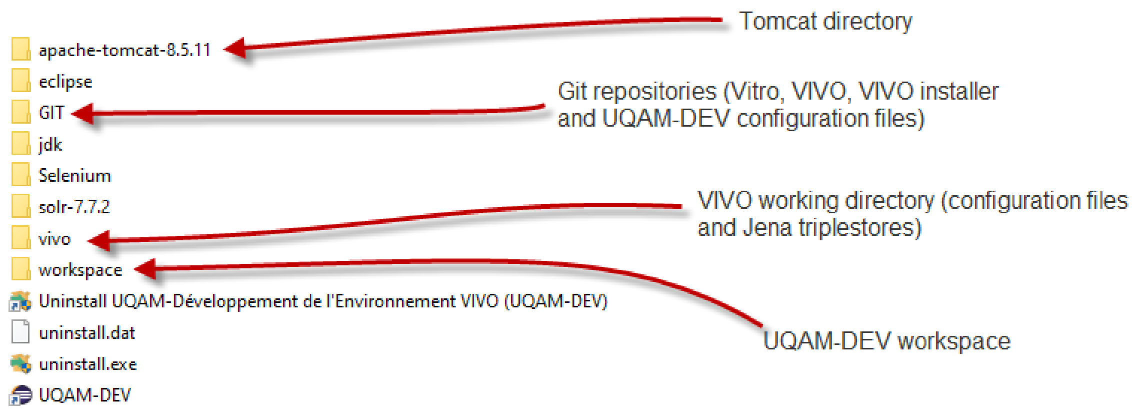

Figure 8: UQAM-DEV Directory Structure

To illustrate its use, we show the standard development cycle of a Semantic Web application (Figure 8), using the VIVO application as an example. We demonstrate how this development process is operationalized in the UQAM-DEV environment.

[13]https://git-scm.com

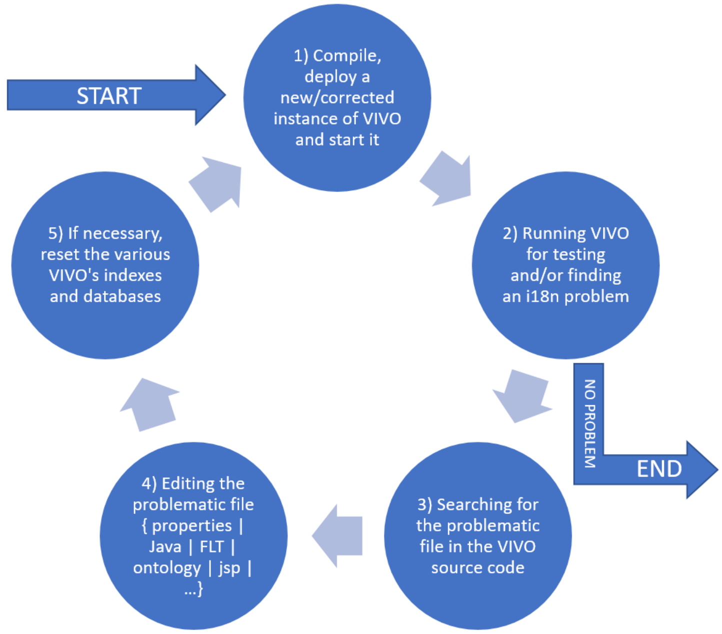

Figure 8: Development Cycle of a Semantic Web Application (VIVO Example)

As shown in Figure 9, the interface consists of a set of visual resources divided into four categories: perspectives (1), menus (2), toolbars (3), and views (4), which are detailed as follows:

1. **Perspectives** allow the user to use the features of the development environment according to the perspective required by his or her role at the time of development. In UQAM-DEV, there are five perspectives: the ontology specialist (Topbraid), the programmer (Java), the tester (Debug), the repository usage (GIT) and the resource usage. Therefore, five perspectives are available.
2. **The menus** correspond to sets of actions that can be undertaken within these perspectives: the menus change according to the perspective chosen.
3. **Toolbars** : adaptable to each perspective, they have icons for the most commonly used functions.
4. **The views:** windows that show parts of the information according to the perspective chosen.

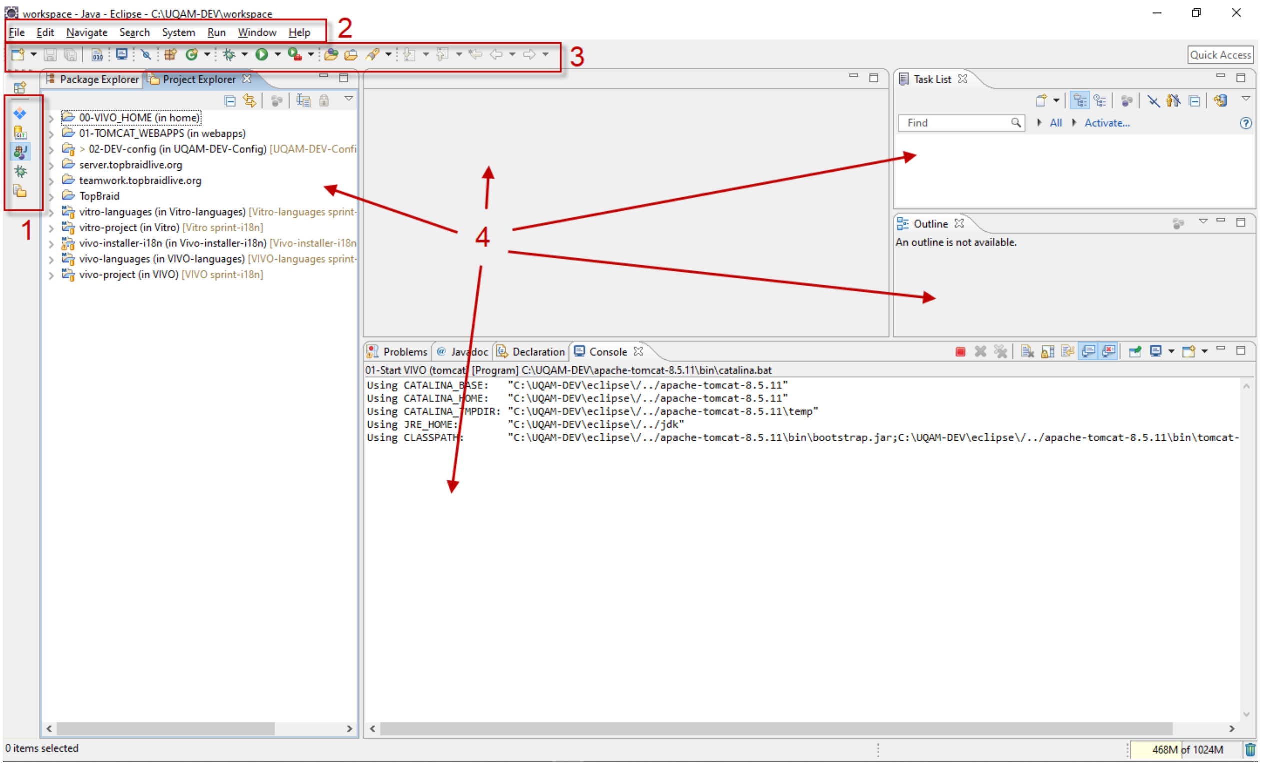

Figure 9: UQAM-DEV Interface

Based on these explanations of the UQAM-DEV functionalities and technologies, on this context (the VIVO development), as well as the interface presentation, we can demonstrate how to obtain the application's code, compile it, install it in Tomcat, and test it. All these steps are performed throughout several exercises in the workshop, so that the participants can try it out firsthand. This is the most hands-on part of the training.

6 - Results, Limitations and Future Developments

This training using a "tiered" educational approach and the outlined learning environment have been successful. After a few days of training, the participants were pleased to have learned a significant amount, while understanding the complexity of the technology. Attendees were very satisfied with the training, which allowed them to acquire rare skills in a relatively short period of time, and the complexity of the technology introduced was presented in a way that was conducive to learning.

Clearly, a few days of training is not enough to master such a complex technology, nor is it a substitute for years of experience. This is the reason why some of the trainees are working on the VIVO project and have the opportunity to use their training and not lose the benefits of these few days of intensive training.

However, there is a need for the participants to further their knowledge. We would like to develop more specialized modules in the coming months, based on target audiences and the skills needed, so that the knowledge acquired during this first training is not lost.

Bibliography

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.